EVALUATION PLAN FOR THE COMPUTERIZED ADAPTIVE [illegible] APTITUDE BATTERY

BERT F. GREEN

C. DARRELL BOCK

LLOYD G. HUMPHREYS

ROBERT L. LINN

MARK D. READASE

DEPARTMENT OF PSYCHOLOGY

THE JOHNS HOPKINS UNIVERSITY

BALTIMORE, MARYLAND 21218

MAY 15, 1982

[illegible] of this plan was sponsored jointly by the [illegible] Research and Development [illegible] Personnel and Training Research [illegible] Psychological Sciences Division [illegible] Office of Naval Research, [illegible] Contract No. N00014-80-C-[illegible] [illegible] Identification Number [illegible]-163.

Report 82-1

Evaluation Plan for the Computerized Adaptive Vocational Aptitude Battery

Bert F. Green, The Johns Hopkins University

R. Darrell Bock, The University of Chicago

Lloyd G. Humphreys, The University of Illinois

Robert L. Linn, The University of Illinois

Mark D. Reckase, The American College Testing Program

Department of Psychology
The Johns Hopkins University
Baltimore, Maryand 21218

May 15, 1982

The development of this plan was sponsored jointly by the Naval Personnel Research and Development Center and by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-80-K-304, Contract Authority Identification Number NR 105-463.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Research Report 82-1	2. GOVT ACCESSION NO.: AD-A115 334
	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): Evaluation Plan for the Computerized Adaptive Vocational Aptitude Battery	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Bert F. Green, R. Darrell Bock, Lloyd G. Humphreys, Robert L. Linn, Mark D. Reckase	8. CONTRACT OR GRANT NUMBER(s): N00014-80-K-304
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, The Johns Hopkins University, Baltimore, Md. 21218	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: PE: 63707N-NPRDC;61153N(42)RR04204 TA: RR0420401 NR: 150-463-1;150-463-2
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs, Office of Naval Research (Code 458), Arlington, VA 22217	12. REPORT DATE: 15 May 1982
	13. NUMBER OF PAGES:
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This study was supported by funds from the Navy Personnel Research and Training Laboratory, and the Office of Naval Research, and monitored by the Office of Naval Research.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Latent trait theory, item response theory, tailored testing, adaptive testing, ASVAB, computerized testing, computerized adaptive testing, ability testing, item characteristic curve theory, evaluation, vocational aptitude battery.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The United States Armed Services are planning to introduce computerized adaptive testing (CAT) into the Armed Services Vocational Aptitude Battery (ASVAB), which is a major part of the present personnel assessment procedures. Adaptive testing will improve efficiency greatly by assessing each candidate's answers as the test progresses and posing items most appropriate for that candidate, thus avoiding items that are too easy or too hard. Computer presentation, recording, and scoring of the ASVAB will improve test security.

(continued on back)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block #20 (Abstract) continued..

This report provides a plan for evaluating proposed procedures for implementing the CAT version of the ASVAB and suggests methods to be used, if CAT is adopted, for checking the utility and operational characteristics of the actual implementation. The report proposes evaluation of item content, dimensionality, reliability, validity, item calibration, item selection and scoring, score equations, human factors. Some special problems include omits, speeded tests, and item bias. Suggestions are also made for the exploring ways of taking advantage of the computerized presentation to get better information from future versions of the ASVAB.

Evaluation Plan for
Computerized Adaptive Vocational Aptitude Testing Program

Contents

Foreword

1 Executive Summary

4 Introduction
6 Armed Forces Selection Tests
11 Adaptive Testing
14 Item Response Theory

24 Components of Evaluation
25 Item Content Specification
26 Dimensionality
32 Reliability and Measurement Error
36 Validity and Differential Prediction
41 Item Parameters - Estimation
47 Item Parameters - Linking
49 Item Pool Characteristics
50 Item Selection and Test Scoring
55 Stopping Rules
56 Equating of Ability Scales
58 Human Factors
61 Special Issues
Speeded Tests
Omits
Item Bias

70 References

85 Distribution List

Foreword

In 1980 the Office of Naval Research and the Navy Personnel Research and Development Center invited a group of experts in psychometrics to review the current plans for implementing a computerized adaptive version of the tests used by the Armed Services for initial personnel selection and placement. That committee consisted of R. Darrell Bock, Bert F. Green (chair), Lloyd Humphreys, Robert L. Linn, and Mark Reckase, with Charles Davis as the ONR monitor. The committee has met with members of CATICC, Computerized Adaptive Testing Interservice Coordinating Committee, and has discussed issues with many other leaders in the field. James McBride, Malcolm Ree, Major Mike Patrow, Hilda Wing and Charles Davis of CATICC have been very helpful. We acknowledge the advice of many colleagues, especially Huyhn Huyhn, Michael Levine, Fred Lord, Melvin Novick, Fumiko Samejima, Hariharan Swaminathan, James Sympson, Vern Urry, and Thomas Warm. This advice was sometimes contradictory, but always helpful. The report that follows is the committee's final recommendation, based on the literature, the advice of others and its own best judgment. The report represents the committee consensus and is to be taken as coming from the committee as a whole.

Bert F. Green
Darrell R. Bock
Lloyd G. Humphreys
Robert L. Linn
Mark D. Reckase

I. Executive Summary

The United States Armed Services are planning to introduce computerized adaptive testing (CAT) into the Armed Services Vocational Aptitude Battery (ASVAB), which is a major part of the present personnel assessment procedures. Adaptive testing should improve efficiency greatly by assessing each candidate's answers as the test progresses and posing items most appropriate for that candidate, thus avoiding items that are too easy or too hard. Computer presentation, recording, and scoring of the ASVAB will improve test security.

This report provides a plan for evaluating proposed procedures for implementing the CAT version of the ASVAB and suggests methods to be used, if CAT is adopted, for checking the utility and operational characteristics of the actual implementation. Suggestions are also made for the exploring ways of taking advantage of the computerized presentation to get better information from future versions of the ASVAB.

The evaluation plan is based on the assumption of a gradual transition to CAT, in which both CAT and traditional paper-and-pencil tests (P&P) will be given. This implies that the CAT version must yield scores that are essentially equivalent to scores from the P&P version. The plan also assumes the availability of prototype adaptive testing equipment before operational implementation.

The role of the computer in adaptive testing is to present each test question (item) on a display screen, to record and score the response, to make new estimates of the candidate's ability after each item response, and to select a next item that will give the best additional information about the candidate's ability. This procedure requires that each test have a large pool of items with widely varying difficulty.

The system's estimates of ability and selections of items are based on a probability model of item responses called item response theory (IRT). The theory provides a curve for each item, showing the probability of a correct answer as a function of ability. In the most widely used model the curve is characterized by three parameters: a, the slope or discriminability; b, the difficulty, and c, the lowest possible probability of a correct answer, called the "pseudochance" level. The CAT procedures discussed here are based on these item parameters.

Specific Recommendations

New items will be needed for the CAT versions of the ASVAB. They should cover the same content as the content in ASVAB forms 11, 12, and 13, which will be the concurrent P&P tests.

Each test in the battery must measure a single ability or dimension. Selecting items that are highly discriminating tends to secure unidimensionality, but this should be verified by other analyses. If more than one dominant factor is involved in a test, proper operation of CAT will require that the test be divided into corresponding subtests.

The standard error of measurement at specified score levels is the best assessment of measurement error in CAT. For some purposes, however, it may be convenient to define an average coefficient of reliability or a reliability at specified score levels. Test reliability should also be assessed empirically by testing a group of examinees twice with different items and correlating the scores.

The validity of the CAT version must be demonstrated. Validity has several aspects. Congruent validity requires that the covariance structure of the tests in the CAT battery match the structure of the standard paper-and-pencil ASVAB, except possibly for scale. Content validity is addressed by item content specification. Empirical validity should be assessed by giving the CAT to persons enrolled in various specialty schools, and obtaining criterion data on their performance in training. To the extent that the CAT battery correlates highly with the paper-and-pencil ASVAB, the validity of the new test battery can be inferred from the established validity of the present ASVAB.

Several competing methods are available for estimating the parameters of each item in each test item pool. The stability and accuracy of estimates obtained by the chosen method should be established empirically by means of simulation studies incorporating realistic error processes. Also, since the parameters for CAT items will have to be determined initially using paper-and-pencil administration, an empirical study is needed to determine what differences, if any, are caused by the different modes of presentation. If it proves necessary to calibrate items in batches, the calibrations will have to be linked. Methods for certifying the linking procedure are proposed. Finally, the characteristics of each item pool must be examined to insure that a reasonable number of highly discrimninating items are available at all ability levels likely to be encountered.

The estimated ability from CAT is on a scale different from that of the P&P tests. A method of equating the scales is required. The details of that method should be reported.

Several possible rules could be used for terminating the testing. Although every test taker could receive the same number of items, theory permits testing until a sufficiently small standard error is achieved. Some compromise rule may be used. The average size of the resulting measurement error must be checked and reported as a function of test score.

The report discusses several human factors in the equipment design and testing procedure, including quiet, glare, legibility, response feedback, and graphics. Immediate display feedback of the chosen alternative is recommended, together with a separate "verify" button to send the results on to the computer. The display screen provides a constraint on item construction: each item must fit on the screen at one time. This constraint may be a problem with some reading comprehension items.

Two of the ASVAB tests, Numerical Operations and Coding Speed, are highly speeded and require special treatment. They will not be adaptive, but will be presented by the computer. The equipment and system design will affect the norming of these tests, so the tests must be calibrated on the operational equipment. Provision for accurate measurement of response times for these items is critical.

On paper-and-pencil tests, candidates may omit items. Experts differ on whether to permit students to skip or omit items in CAT. It is recommended that omits not be permitted.

It is important that the test scores and the items not be biassed in favor of any subgroups of persons. Studies are recommended of potential differential validity of tests for men and women, and for various ethnic groups. Studies of item bias are also recommended. Since similar studies have recently been made for the current ASVAB, these studies can await implementation of the system. The nature of the computer presentation itself is not expected to favor any one group.

General Comments and Recommendations.

In general, the procedures used in CAT should be thoroughly documented and explained.

Apart from the specific projects proposed to evaluate the CAT version of the ASVAB, some research projects should be undertaken or supported to improve aspects of the procedures. CAT methods are still under development and further development is needed. Multidimensional models, and models that analyze response option characteristics should be developed further.

Other ways of using the computerized testing equipment should be explored to get more information from the ASVAB, and eventually to alter the ASVAB by including new kinds of measures. Such studies will provide an additional return on the investment in CAT.

II. Introduction

The United States Armed Services are currently considering the introduction of computer-based technology into procedures for evaluating the cognitive abilities of personnel. Computer presentation, recording, and scoring of standardized tests can be expected to make the tests more secure and the testing more efficient. With computer presentation, the test can be adapted to each candidate's level of ability, assessing each candidate's answers as the test progresses and selecting the items that will give most additional information about the candidate's ability. Adaptive computer presentation promises the same accuracy of measurement as present paper-and-pencil (P&P) tests in much less time. Although present plans are to devise computer implemented versions of the existing Armed Services Vocational Aptitude Battery (ASVAB), computer technology has the potential for a wider array of assessment measures and will ultimately provide more effective assignment of personnel.

The introduction of computerized adaptive testing (CAT) in the U.S. military enlistment procedures would be the first large-scale operational use of this technology. It is therefore especially important that the proposed procedure be carefully and thoroughly evaluated. Such an evaluation can facilitate a fully informed decision about whether to proceed with the operational implementation. If the decision is positive, as we expect, then the operational use of CAT should be systematically monitored and evaluated, to insure that the new method is working as expected. Also, further research and development should be done to enable the Armed Forces to get the maximum benefit from the system.

Computer methods represent a large change in personnel testing. Methods of evaluating tests must be revised to suit the new procedures. The concept of validity is still central, but the concepts of reliability and scale equivalence take on new meanings and must be evaluated differently. New issues arise in reporting the efficiency and dependability of an ability test. Consequently the evaluation of a computerized version of the test must be formulated in its own right, and not merely taken over from traditional test development practice.

This report is concerned with the evaluation of a computerized adaptive version of a particular test battery, the Armed Services Vocational Aptitude Battery (ASVAB). The report discusses the empirical evidence that will be necessary and/or highly desirable for establishing the psychometric suitability of the new version. It considers each of the major psychometric properties - dimensionality, reliability, validity, and score calibration - as well as special problems unique to a computerized adaptive test - item calibration, adaptive selection of items, scoring the test, and the human factors that affect the use of the computer equipment. Other problems also addressed are item bias, speededness, and the calibration of new items in an operational context. Finally, opportunities for further gains through computer methods of testing are suggested as

topics for further research and development.

Some of the studies proposed are already underway and some of the procedures endorsed are those that have already been selected. In this document we make no distintion between what should be done and what already is being done. Rather, we report on the entire range of psychometric problems in adaptive testing. Also, recommended methods are based on current knowledge and may be superceded by new developments.

The committee anticipates that the feasibility of CAT will be established, and that CAT will be available for implementation. CAT is expected to provide more efficient use of available testing time and improved test activity. While these are the immediate economic benefits, CAT also has the potential for many improvements in personnel selection and placement procedures. The committee urges that the potential economic benefit of future capabilities also be considered, and that additional work be supported to develop some of the many possibilities.

Armed Forces Selection Tests

The Armed Services of the United States use standardized tests of skill and knowledge as part of their personnel recruitment and placement procedures. The Army General Classification Tests (AGCT) were used extensively during the second world war. In 1950, the Armed Forces Qualification Test (AFQT) was introduced as a replacement for the wartime tests. The AFQT scale was calibrated to the AGCT scale using a 1944 wartime reference group, although the test was somewhat changed in content.

New forms of the AFQT were introduced in 1953, 1956, and 1960. Starting in 1972 and continuing through 1975 each service used its own test battery. However, each battery provided an estimate of an equivalent AFQT score, as well as other scores. In 1975, the Services again began to coordinate their selection testing efforts, resulting in a new, expanded test battery, called the Armed Services Vocational Aptitude Battery (ASVAB). An early version of this battery had been in use in a high school recruitment program. New forms were developed, and some content changes were made for the new forms, ASVAB 5, 6, and 7, which were to be used as the standard military recruitment tests. Again the scale for the parts of the ASVAB making up the new AFQT composite was calibrated to the earlier AFQT scale, although content by now was considerably expanded from the original tests.

A calibration problem arose with forms 6, and 7 of the ASVAB that resulted in too many underqualified persons being accepted into the service. The calibration was studied by several groups within the Department of Defense, and studies by outside groups were also commissioned. Each study made a slightly different recommendation for change. A special outside technical committee (consisting of Robert Linn, chair, Melvin Novick and Richard Jaeger) was appointed to evaluate the studies. They recommended a change in the calibration table that solved the problem. (Jaeger et al, 1980) This episode is mentioned to emphasize that calibration is a critical aspect of any form of the ASVAB.

In 1978, a study was made of the possibility of applying to the ASVAB the growing technology in computerized adaptive tests (CAT). Computer presentation has the potential advantage of improved test security, as well as simplifying test scoring and reporting. Mainly, though, computer presentation permits adaptive testing, or "tailored" testing as it is sometimes called, which can reduce testing time by using testing time more efficiently. The Computerized Adaptive Testing Interservice Coordinating Committee (CATICC) was formed to plan for development and implementation of a CAT version of ASVAB.

At about the same time, the ASVAB was being slightly restructured, and new forms were being developed. ASVAB forms 8A, 8B, 9A, 9B, 10A, and 10B became operational in October 1980. New forms of the same test, with no

Table I. Tests of the ASVAB Forms 8, 9, and 10

Tests included in the Armed Forces Qualifying Test composite are enclosed in dotted lines.

Tests	Number of Items		Testing Time (minutes)
1. General Science	25		11
2. Arithmetic Reasoning	30		36
3. Word Knowledge	35	AFQT	11
4. Paragraph Comprehension	15		13
5. Numerical Operations (speeded)	50		3
6. Coding Speed (speeded)	84		7
7. Auto and Shop Information*	25		11
8. Mathematics Knowledge	25		24
9. Mechanical Comprehension*	25		19
10. Electronics Information*	20		9
Total Questions	334		
Total Testing Time:			2 hrs. 24 mins.

*Some test items include diagrams.

content change, are now being developed for introduction late in 1983. These new forms are ASVAB 11A, 11B, 12A, 12B, 13A, 13B. If the computer version of the ASVAB is approved, it could be implemented late in 1984, but would be phased in gradually. Thus the CAT version and the new paper-and-pencil versions of the ASVAB would have to be as nearly equivalent as possible, although each may have its respective norm table, a candidate should have the same probability of selection and classification no matter which version of the test is taken.

Currently the ASVAB consists of the ten tests listed in Table I. Numbers of items and testing times are indicated. These times do not include the time to read and understand the directions, nor the time for any other administrative details, including rest periods. When all these things are taken into consideration, the ASVAB takes about 4 hours to administer. (McBride, private communication).

Except for Tests 5 and 6, which are speeded, the timing has been established so that most test takers will finish most of the tests. Data show that, excluding the two speeded tests, each item is attempted by about 98% of the test takers, on the average. The last item on each test is attempted by 92.2% of the test takers on the average. The raw score on the test is the number of correct answers. There is no penalty for guessing.

The instructions for the ASVAB say, "Remember, there is only ONE BEST ANSWER for each question. If you are not sure of the answer, make the BEST GUESS you can." Each test includes the instruction, "Don't spend too much time on any one question." (On a recent survey of a national probability sample of ASVAB takers, about 2/3 said that they did guess, the others said they did not.)

The use of long time limits on all but the speeded tests makes the ASVAB a power test, which is good measurement practice when speed is not being explicitly evaluated. But long time limits raise administrative problems, since many test takers finish a test long before the time limit, and are forced to wait idly, with possible adverse effects on anxiety and motivation. There would be no such waiting with a CAT version, because each candidate proceeds at his or her own pace.

The raw score on the AFQT is a composite of the raw scores on Tests 2, 3, 4, and 5, with Tests 2, 3, and 4 getting unit weight, and Test 5 getting 1/2 weight. Various branches of the Armed Services use other composites of their own design for selecting applicants to the Service and for selecting applicants to clusters of specialty schools (Maier & Grafton, 1981b).

ASVAB scores are used to make two kinds of decisions. First the scores are used to decide if the candidate is qualified to enlist in his or her chosen service. At present this decision is based on the AFQT score, a composite of four of the ASVAB test scores, as described above. However, each service has a different cut-off score to determine qualification. None of the services admit persons who are in the lowest 9% of the reference population (1944 AGCT test takers) and many of those in the next

10% are rejected as well.

A second type of decision that depends on the ASVAB is whether the person is qualified for a particular specialty school, or a particular set of specialty courses. Each specialty school, and sometimes each particular course has its own entrance criterion, based on a particular combination of test scores, with a particular cut-off. There are literally hundreds of different specialties, with different composites and cut-offs. Of course, admission to certain advanced schools requires not only certain test scores but also successful completion of earlier training programs.

With decisions being made in so many diverse ways, it is not possible to focus attention too closely on any one test score. Even the first level decision is based on a composite of several test scores, the AFQT. It will therefore be necessary to provide scores that have good accuracy at all score levels.

The two speeded tests are Numerical Operations and Coding Speed. Special methods are needed for computer versions of these tests. The precise nature of the response, i.e. the human factors component, is critical in these tests.

The Navy Personnel Research and Development Center (NPRDC) now has an experimental installation where computerized tests are given to Armed Forces personnel, as a part of test research and development. In addition, a new experimental facilty is being set up with an array of computer-driven terminals to test computerized version of the ASVAB as they become available. NPRDC plans to begin preliminary evaluation of a CAT battery in 1982. Three of the ASVAB tests use elaborate diagrams and drawings: Auto and Shop Information, Mechanical Comprehension, and Electronics Information. Although the current experimental facility at NPRDC does not now have the capability for graphical items, the prototype system and operational systems are intended to include a graphics capability.

The Air Force Human Resources Laboratory has contracted for the development, pretesting and statistical analysis of an experimental pool of 200 items for each of the ten ASVAB tests. In addition, operational item pools are currently being developed. Plans call for 200 items in each operational test pool.

Some evaluative work on CAT is already under way. Some of the studies recommended in the present report are already planned or are in progress. The present report may lead to some change of detail in those studies, and in any event puts them in the context of the over-all evaluation, and supports their execution.

One large-scale study of the current paper-and-pencil ASVAB has recently been completed; reports of data analyses will be issued as soon as possible. The Profile of American Youth is a national longitudinal study, in which a carefully designed sample of persons in the United States age 16-23, took form 8A of the ASVAB, so that national norms could be constructed. Considerable additional information about the ASVAB can be obtained from this excellent data base (Department of Defense, 1982).

The ASVAB technical manual (Wilfong, 1980) contains much useful information. Past history was culled from a review by the ASVAB Working Group (1980). See also Department of Defense, 1980.

Adaptive Testing

The principal idea of adaptive testing is simply that each test taker is asked questions that are appropriate for his or her level of skill or ability. It is inefficient to ask questions that are too easy or too difficult for the candidate, since those responses contribute very little information about that person's ability. The terms adaptive testing and tailored testing will be used as synonyms in this report.

The method of adaptive testing has roots in early psychological measurement. Psychophysicists, beginning with Wundt, determined sensory thresholds by presenting stimuli at varying intensities according to the observer's ability to sense them. Binet, (1909), the father of mental testing, asked each child questions appropriate to the child's age, and moved up or down the age scale depending on the child's answers. The process of choosing items appropriate to the child's mental ability can be viewed as fitting the test to the test-taker, hence the term tailored testing. Such a procedure is very difficult to manage if people are tested in groups rather than one at a time, so ordinary pencil-and-paper (P&P) tests present the same items to all test-takers. The items on group tests vary in difficulty over a range appropriate to the population being tested, so group tests are roughly matched to the population, but cannot be tailored to the individuals.

Several attempts have been made to approximate a tailored test using a P&P mode. Lord (1971) proposed a flexilevel test in which items were ordered in difficulty. Everyone started with the item of median difficulty. A special answer sheet revealed whether a response was correct or incorrect. Whenever a candidate answered an item correctly he tried the next harder item that he had not already tried, and whenever he got an item wrong, he tried the next easier one.

Another procedure consists of a routing test followed by a second test selected from a series of tests that are graded in difficulty (Lord, 1971). The score on the routing test indicates which second test a candidate is to take. Both schemes are more efficient than a conventional test. But both schemes are cumbersome, and neither gains the full efficiency possible with an individually tailored test. An experimental comparison of these procedures has been made by Friedman, Steinberg, and Ree (1981).

With a digital computer to present the test items, item-by-item adaptive testing becomes feasible. The computer can score each response immediately and can then select the next item that will be most appropriate for the candidate. Each candidate gets a set of items uniquely selected for him or her. More specifically, each person's first item generally has about medium difficulty for the total population. Those who answer correctly generally get a harder item; those who answer incorrectly get an easier item. After each response, the examinee's ability is estimated, along with an indication of the accuracy of the estimate. The next item to be posed is one that will be especially informative for a person of the estimated ability, which generally means an item for which the probability

of a correct response, at that ability level, is in the neighborhood of .65. Normally, the process results in harder questions being posed after correct answers and easier questions after incorrect answers. Ideally, the change in item difficulty from step to step is usually larger earlier in the sequence when less is known about candidate's ability, but later in the sequence the difficulty changes less radically as the system tries to refine its estimate of the candidate's ability. The process continues, until there is enough information to place the person on the ability scale with a specified level of accuracy, or until some more pragmatic criterion is achieved. If desired, each candidate's score on a CAT can be estimated to the same level of accuracy. By contrast, high and low scores on a group test are typically less accurate than scores near the mean.

A CAT consists of a set of items, called an item pool or item bank, from which particular items are selected for presentation to the candidate. The precision of the CAT depends on the characteristics of the items in the pool. If the pool is not large enough, and is not well-matched to the ability distribution of the group being tested, the advantages of an adaptive test will not be fully realized. If for example, the adaptive procedure indicates that the next item for a particular person should be moderately easy, but there are no more moderately easy items, the system will have to settle for an item that is very easy, or for one that is moderately difficult, with the result that less information will be obtained than if an appropriate item has been available. Thus adaptive testing requires a sufficient supply of items at each ability level. If security considerations suggest that the items be varied, this implies a need for several alternatives at each ability level, so large item pools are needed for adaptive tests.

Adaptive testing places new demands on psychometric test theory and method. Classical test theory is not adequate; methods appropriate for group tests will not work with adaptive tests. The most obvious problem is that the test score can no longer be the number of items answered correctly. In an ideal tailored test, after the first few items, everyone will tend to answer about the same number of items correctly. The score must depend in some way on the characteristics of the items answered correctly.

Also the indices commonly used to judge the quality of the items are less appropriate. The standard index of item difficulty is the proportion of persons answering the item correctly, which is dependent on the population of test takers. Likewise, the standard indices of item discriminating power, such as the item-test correlation, are also dependent on the population.

Finally, adaptive tests place more stingent demands on the test items in the item pool. Adaptive tests are presently designed to work with items all measuring a single aspect or dimension of ability. Adaptive testing is based on the notion of items and people placed along a single scale of ability. Unidimensionality of the test items is therefore central. Although adaptive methods may eventually be developed for multidimensional test domains, present procedures expect a single dimension. When a test

has one strong dimension, but several facets, as when verbal skill is measured by different types of items (antonyms, analogies, etc.), then special precautions are needed in an adaptive environment to balance the facets. This issue is discussed in more detail in the body of the report.

Some concern has been expressed about possible legal challenges to the equity of adaptive testing. The fact that the candidates do not take the same items might be interpreted to mean that they do not all take the same test. CAT might be challenged for not permitting some candidates to display their ability, because it does not give them the opportunity to answer the more difficult items. Such a challenge may possibly be raised, but it seems to us to be without merit. At present, not all candidates take the same P&P test form. In most testing programs there are several different test forms, all calibrated so as to be equivalent. The questions differ, but the area of skill or knowledge assessed is the same on all test forms, and every candidate has the same opportunity. In the same way, every candidate's encounter with the CAT form of the test offers the equivalent opportunity. Indeed one of the overriding considerations in the evaluation for CAT recommended in the present report is the assurance of equivalence, so that each candidate does have the same fair chance.

It should be noted that the concept of fairness involves equal opportunity, not equal treatment. In a track-and-field meet, each competitor must have the same chance at the high jump, but fairness does not require that a person who can't clear a six-foot high jump nevertheless be given a chance at seven feet. The point is to see how high each person can jump, not to permit each person license to try all levels. In a tennis tournament, it is not considered necessary for every player to play the best players - only that every player have the same initial chance. In the same way, a CAT provides every candidate the same initial opportunity. Further, those who fail the first two or three items can still get a good score if they pass all the subsequent items. CAT continually gives each candidate additional chances. No fairness is lost by not asking a candidate questions that are too easy or too difficult. Indeed, by providing more accuracy for high and low scores, the test is potentially more fair.

Early work on adaptive testing is discussed in Harman et al, (1968); Holtzman (1970), and Wood (1973). More recent accounts can be found in U.S. Civil Service Commission (1976), and Weiss (1974, 1978, 1980). Applications have been discussed by Urry (1977), Lord (1977a,b), and Kreitzberg & Jones (1980).

Item Response Theory

Classical test theory is not suited to adaptive tests. Classical theory supposes that all test-takers confront the same set of test items, as in the conventional P&P tests. Classical indices of reliability, validity, and item quality are relevant to a particular set of items and a particular population of test-takers. But an adaptive test is different for each taker, and is, in principal, independent of the particular population.

A theory that is appropriate for adaptive tests was developed by Rasch (1960), Lawley (1943), Tucker (1946), Lord (1952), Samejima (1969), Owen (1975), and others. This new theory, now called item response theory (IRT), was discussed by Birnbaum (1958) as latent trait theory, and appears in Lord & Novick's (1968) major treatise on test theory. Hambleton & Cook (1977), and Warm (1978) give good introductions. More complete accounts of IRT have been given recently by Lord (1980), Urry & Dorans (1980), and Urry (1981).*

The theory postulates that persons vary in the ability being assessed by the test, and that their abilities are distributed along a continuum labelled θ, from low to high. Each person has a particular ability level; the ability of Person i is θ_i. The probability of answering an item correctly is assumed to vary with ability, symbolized for Item j by $P_j(\theta)$. The model assumes a particular form for this probability function. The traditional choice is the cumulative normal function (ogive) but the cumulative logistic curve is essentially indistinguishable from the cumulative normal, and is mathematically convenient. Its mathematical form, shown by Items 1 and 2 in Fig. 1, is

$$P_j(\theta_i)=1/[1+e^{-u_{ji}}]$$

where

$$u_{ji}=1.7a_j(\theta_i-b_j).$$

---

* The term "latent trait theory" is used in the earlier literature, rather than "item response theory." "Latent" signifies that the ability or skill being assessed is inferred from the item responses, and is in this sense latent in the item responses; "trait" merely refers to a characteristic of the examinee that is sufficiently stable to be measured. However, some laypersons may interpret the terms "latent trait" in a non-technical sense as implying a fixed, inherited property of the individual not alterable by training. This interpretation is incorrect, and is in no way appropriate to tests of vocational skills and knowledge, so the neutral phrase "item response theory" is preferred.

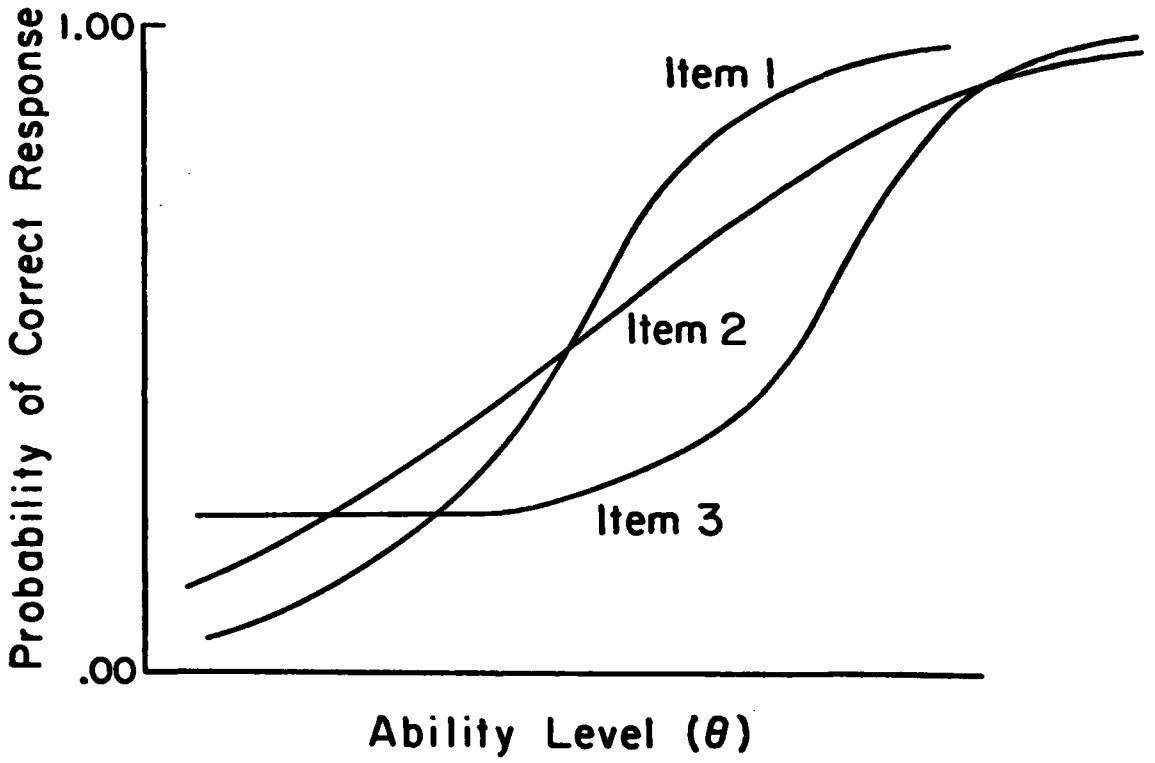

Figure 1. Illustrative Item Response Curves. Item 1 is more discriminating than Item 2 Item 3 includes guessing.

So far, the model is not novel, but is simply borrowed from psychophysics, where $P_i(\theta)$ is the probability of detecting the presence of some stimulus, or from biological assay, where $P_i(\theta)$ is the proportion of samples that exhibit some property. What is special about the test theory context is that the ability values, the θ_i's, are unobserved, and indeed are unobservable. The nature of the θ variable is in part determined by the assumption that the response curve of each item has the form of Equation (1), varying only in a_j and b_j. The nature of θ is further specified by the fundamental assumption of the model that, for a fixed value of θ, responses to the items are independent. Thus, the probability that a person of ability θ answers both items j and k correctly is simply

$$P_j(\theta) \times P_k(\theta).$$

This is called the assumption of local independence. It means, in essence, that the item responses are related to each other only because they are all related to the ability scale, θ. The source of the interitem relationship is the underlying ability θ, which is being measured by the items. This assumption is fundamental to many models of individual differences, including common factor analysis and latent structure analysis.

The model described above is frequently called the two-parameter model because each item response curve has two parameters, a_i and b_i. A simpler, one-parameter version of the model has many attractive features; it is obtained by assuming that all items can be treated as being equally discriminating, so that $a_j = a$ for all j. The resulting model, also called the Rasch model, has been advocated by Andersen (1973), Fischer (1973), and Wright (1977). Unfortunately this model does not fit most data. Items are not equally discriminating, and the inequality matters. Koch & Reckase, (1978), and Patience & Reckase (1979) showed that the more complicated models performed better than the Rasch model. The simple model does not take differences in item discrimination into account when selecting items to present.

Neither model is adequate for multiple-choice items, in which the item may be answered correctly by chance. Some test-takers guess when they don't know the right answer, and sometimes they are lucky. Because of guessing, the probability of correctly answering the item does not necessarily decrease to zero for persons of very low ability, but may decrease to some minimum level, often called the pseudo-chance level. The pseudo-chance level for an item becomes the third parameter of the item, c_j.

In the three-parameter model, the logistic item response curve, indicating the probability of a correct response to item j, becomes

$$P_j(\theta)=c_j+(1-c_j)/[1+e^{-u_{ji}}].$$

Where as before,

$$u_{ji}=1.7a_j(\theta_i-b_j).$$

Item 3, in Figure 1 has such a response curve. For very low values of θ,

$$P_j(\theta)=c_j.$$

As θ increases, the probability rises from c_j to 1, in the same way that it rose from 0 to 1 in the earlier model that does not include the c_j parameter.

It might be supposed that for four-option items like those on many tests, c_j would be about .25. However, it is often found that c_j is less than would logically be expected if wrong answers were random guesses. Not all examinees guess when they do not know the correct answer, and wrong answers may be due more to misinformation or incomplete information than to guessing. One study shows that on some 4-alternative multiple choice tests, the c parameter varies from .10 to .35 or more, with a median of about .20 to .25. Another study finds that if all item response curves for a similar test are forced to have the same c value, a value of .10 is best (Bock & Mislevy, 1981). Adding the third parameter, c_j, complicates item response theory enormously, and it would be an immense convenience to leave it out. Nevertheless, the three-parameter model <u>is</u> needed. The model does not fit multiple choice items well when $c_j = 0$.

The classical theory of statistical estimation provides a powerful way of describing the amount of information in an item, and in a test. Test information is inversely related to the variance of measurement error. (Bayesian theory provides an equivalent result.) The relative amount of information that an item provides about persons of various abilities is called the item information function. It is given by

$$I_j(\theta_i)=[P'_j(\theta_i)]^2/[P_j(\theta_i)\cdot Q_j(\theta_i)]$$

where

$$Q_j(\theta_i)=1-P_j(\theta_i)$$

and

$$P'_j(\theta_i)=\frac{d}{d\theta}[P_j(\theta_i)].$$

It can be shown that maximum information occurs where the curve is steepest, which is in the vicinity of $\theta=b$; and that this information is proportional to a. This means, first, that a is indeed an index of discrimination, and second, that it is best to use items with b-values near to a person's ability. (The specific location of the maximum, and the specific information at that point depend in a complex way on c.)

For a test of fixed items, the information function of the test is simply the sum of the information functions of the individual items. It is easy to see that for a fixed-item test to yield a reasonable amount of information about persons of a wide range of ability, the item difficulties must span a comparable range, with the result of not providing very much information anywhere. In practice a compromise is usually struck. Tests are constructed with not many very easy items, nor many very hard items. The information curve for a test of Arithmetic Reasoning, shown in Figure 2, is typical of many standard tests. The height of the test information function shows the relative precision with which test scores are measured. Figure 2 shows that low scores and high scores are not measured very precisely. The reciprocal square root of the test information function is asymptotically proportional to the width of the confidence interval for

estimating θ from the item responses, which in turn is proportional to the standard deviation of the measurement errors for each fixed true ability level.

The relationship between ability defined by the item response model and the number right score is in general non-linear. Figure 3 sketches the relation of the ability scale, θ, and the expected number-right score for a typical paper-and-pencil test. The relationship is nearly linear in the middle of the range, but is curvilinear at the extremes.

For an adaptive test, we can characterize the entire pool of available items by the information function of the pool, which is the sum of the individual item information functions for all the items. This information function would be a much broader function than that in Figure 2. But the important issue is the information function for the items given to a particular candidate. In adaptive testing, the tailoring process chooses from this pool so that the information function for each candidate will be maximum in the vicinity of that candidate's ability level, θ_i. Thus, with an adaptive test, either the precision of measurement is greatly improved, if the number of items is not changed, or a given level of precision can be achieved with a much smaller number of items than would be possible with a standard fixed-item test. (Bayesian theory provides a slightly different analysis but reaches the same conclusions.)

It is important to recall that at the start of the testing process we know little or nothing about the candidate's ability level. Consequently, in a tailored test, the first item presented is one that is appropriate for the average candidate. (Performance on any previous test in the battery may be used to improve the initial choice.) After each item response, an improved estimate can be made of the candidate's ability, and more appropriate items selected for presentation. At each stage of the process we have not only an estimate of the ability of the candidate but also an estimate of the standard error of the estimate so we know how good our current estimate is. We may stop when this confidence interval becomes narrow enough, or we can stop after a fixed number of items, chosen so that, on the average, the level of precision is acceptable.

In adaptive testing, the estimate of ability and the choice of the next item require knowledge of the parameters of the item response curves - the a's, b's. and c's. Estimates of these values must have been determined before the testing process is begun. This is usually done by giving all of the items to comparable, large samples of candidates, in a

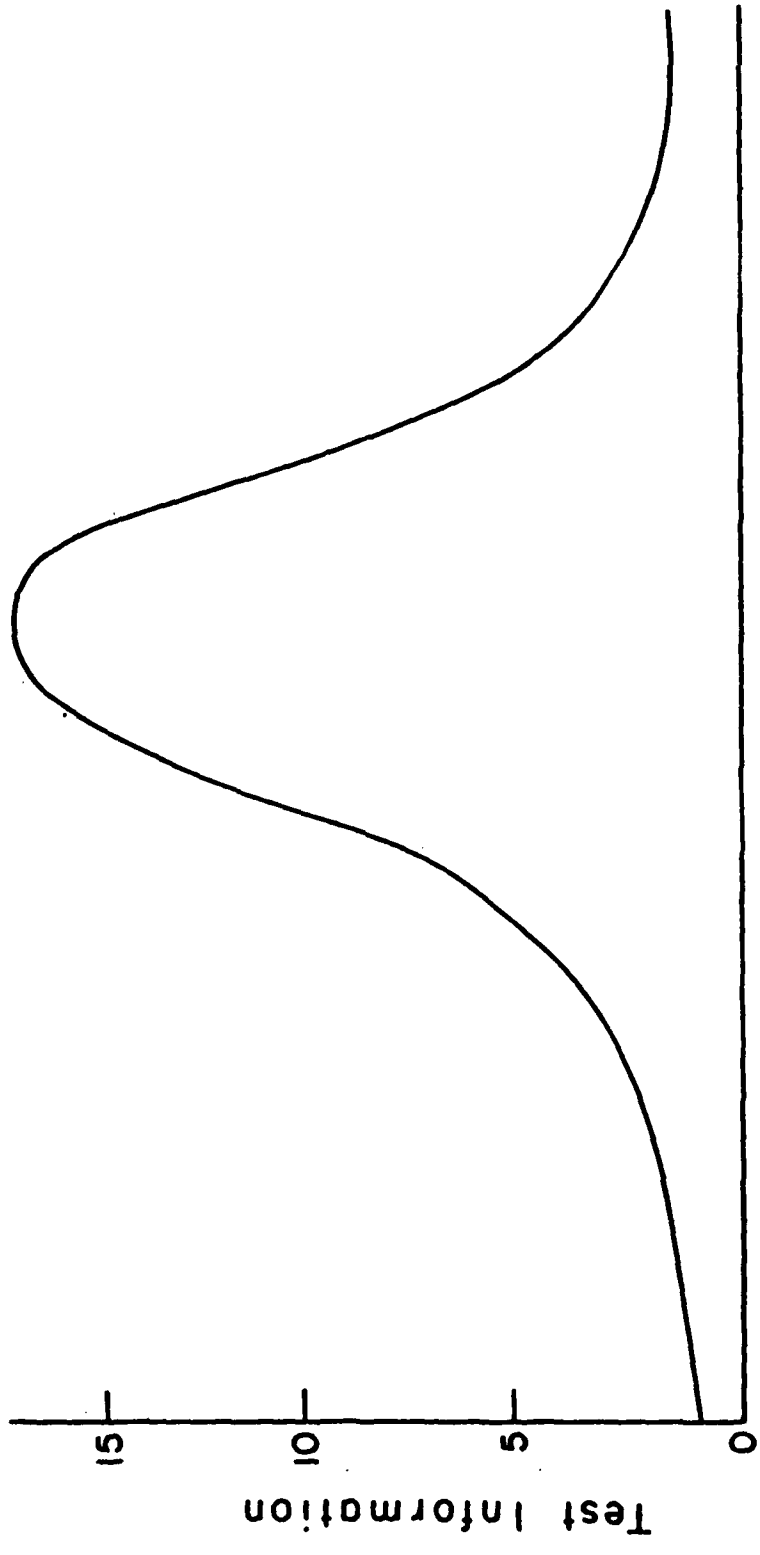

Figure 2. Test Information as a Function of Test Score on a Representative Conventional Test.

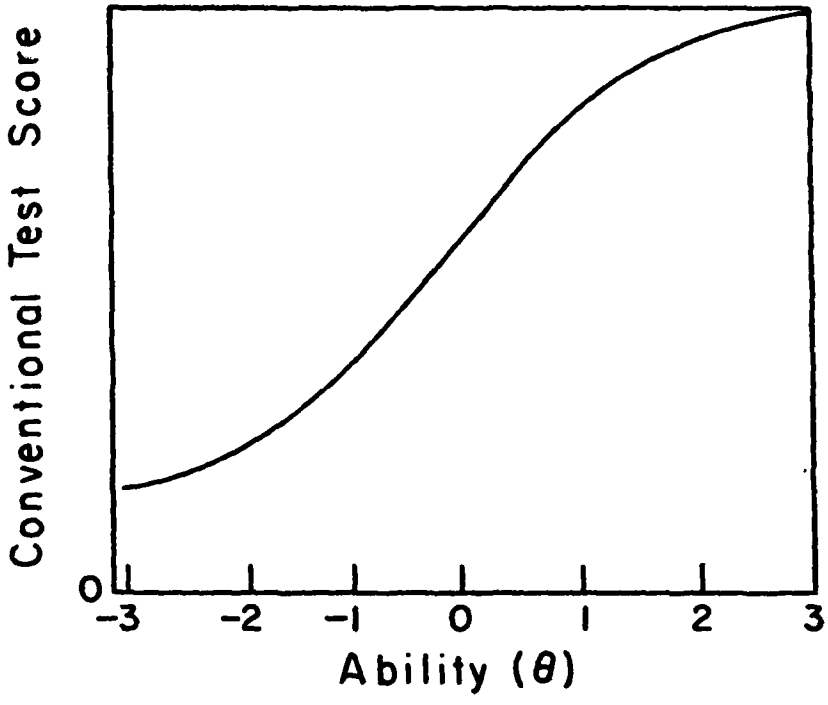

Figure 3. Relation of Ability Scale (θ) to the Number-right Score on a Conventional Test.

standard testing situation. If there are too many items for this to be practical, then overlapping subsets of items can be given to several different samples of candidates. Methods are then available for linking the estimates of item parameters. There is a large literature on parameter estimation; see for example Reckase (1978), Ree, (1981) and Yen (1981).

The step of determining the item parameters in advance is also a part of conventional testing, where item difficulty and item discrimination indices are obtained from pretest data. But these values are used in a somewhat informal way in constructing a conventional test, whereas the item parameters are a central part of the adaptive testing process.

Experts agree on the general outline of the above process, but disagree about details. Most experts advocate using the three-parameter model for multiple choice tests, although some advocate the one-parameter logistic model (Wright, 1977). Some experts advocate using "maximum likelihood" estimators of ability, and corresponding estimates of item parameters. Lord (1980) and Samejima (1977a,b) are the chief advocates of this position; Wood, Wingersky & Lord (1976) have authored a widely used computer program, LOGIST, to compute item parameters. Others, including Urry & Dorans (1980), advocate Bayesian methods, in which an initial prior estimate of ability is made, together with a guess about the ability distribution, and the item response data are used to improve the initial estimates by Bayes's theorem. Urry (1981) has prepared computer programs for item estimates called OGIVIA and ANCILLES, to provide Bayesian estimates of item parameters. McKinley & Reckase (1980, 1981a) give a comparison of ANCILLES and LOGIST.) Bock & Aitken (1981) advocate a marginal maximum likelihood approach to estimating item parameters, and Bock & Mislevy (1981) offer a program called BILOG based on this method. Bock advocates empirical Bayes estimation of ability in the tailored testing situation, possibly with reduced weighting of extreme responses.

The experts also disagree on the best procedures for selecting successive items in the tailoring process, and in the criterion for stopping the process. From a purely theoretical perspective, the next item to be given to a person should be the most informative item, as judged by the current estimate of the person's ability. Slavish following of that rule is likely to result in a few of the very best items - items with the largest a_j's being used a great deal, with other items in the pool possibly under-used, which may jeopardize test security. In practice, there will usually be many items that are almost as good as the best in any situation, and it may make very little practical difference which one is selected from among these possible items. NPRDC is now conducting a series of computer simulations of adaptive testing to evaluate the psychometric effects of alternative adaptive testing procedures, including random choice of test items in the vicinity of the current ability estimate of a candidate.

Also, most experts advocate continuing the testing process until a predetermined level of accuracy of the test score (the estimated ability) is reached. Others feel that little may be lost in practice if the same number of items is given to each test taker. A compromise may be to aim for a predetermined accuracy, but to place an upper limit on the number of items to be given.

In a field as new as computerized adaptive testing, there are sure to be disagreements among experts. The committee members themselves are not in complete agreement about all aspects of the evaluation. Perhaps it would be more nearly accurate to say that there are several issues about which there simply is not yet enough information for a decisive answer. But the committee is in complete agreement that the proposed procedures are satisfactory, and that the computerized version of the ASVAB can be as good as the present version, if not better. The proposed studies will indicate whether in fact the computerized version does live up to its promise. We expect that it will, possibly with some adjustments. Undoubtedly there will be room for improvement, but we are not now able to foresee the potential improvements. We do not believe that further theoretical developments will resolve these issues, most of which are at the interface between theory and practice. Current development is at a stage where implementation is needed in order to obtain further information about many aspects of the procedures.

III. Components of Evaluation

This section lists proposals for evaluating the various important properties of a CAT, including unidimensionality, reliability, validity, and equivalence among test forms. Since all of these properties depend critically on the way in which the test is implemented, proposals are also made for evaluating the quality of the procedures for determining item parameters, the procedure for tailoring the test to the individual through item selection; and the procedure for determining the final test score. Some important aspects of human factors in the equipment for adaptive testing are noted. Some special problems are discussed, including the speeded tests, the question of whether omitting is to be allowed, and item bias.

Throughout we have adopted the style of the Standards for Educational and Psychological Tests published by the American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education (APA, 1974). Recommendations are stated succinctly, and are rated "essential", "very desirable", or "desirable." Discussion accompanies each recommendation.

In this report, the terms "we", "us", and "the committee" refer to the five authors, or a substantial majority of the authors. The following abbreviations are used throughout:

ASVAB	Armed Services Vocational Aptitude Battery
CAT	Computerized Adaptive Test
CATICC	Computerized Adaptive Test Interservice Coordinating Committee
CEPCAT	Committee for an Evaluation Plan for Computerized Adaptive Tests
IRT	Item Response Theory (also called Latent Trait Theory)
IRC	Item Response Curve (also called item characteristic curve, item operating characteristic, and item response function)
NPRDC	Naval Personnel Research & Development Center
P&P	Paper and Pencil (conventional group test)

Item Content Specification.

In a sense, the specification of the item content is not within the areas covered in this report, which is focused on technical issues. Still, content is of fundamental importance. At present, it is important that the CAT items cover the same ground as the P&P tests with which they are intended to be interchangeable.

C1. Specifications for item content should be the same for both the CAT and P&P ASVAB tests. Essential.

Beyond this obvious requirement, there is nothing special about CAT items that does not apply equally to P&P items on conventional tests. The purpose of this requirement is to help insure the comparability of CAT and P&P forms of the ASVAB, on the assumption that for an initial period, both test modes will have to be used, while CAT is being introduced. When CAT is established as the primary testing mode, with P&P versions used rarely, if at all, then this requirement is withdrawn. Indeed, we urge using the new medium to improve assessments through new and expanded content.

Frequently there are informal guidelines about coverage or other characteristics of items for a given test. Wherever possible those guidelines should be made explicit, so that they are clearly understood by those preparing the items.

Elsewhere it is recommended that items be selected that are highly discriminating (i.e., have high values of a_j.) Assuming that such a criterion is followed it will be important to examine the selected items for coverage of content, to be sure that item selection has not disturbed content specifications.

C2. The content of items selected for the final item pool should match the content specifications. Essential.

C3. Test items must be compatible with CAT equipment. Essential.

Test items must be designed to be consistent with whatever equipment is to be used. At present, the main constraint is that each item must fit entirely on the display screen at one time. This can be a problem for reading comprehension items, which typically involve a paragraph to be read, and one or more questions to be answered about the paragraph.

Dimensionality

Present methods of adaptive testing, and the item response theory (IRT) on which the methods are based, require that the test be unidimensional. Each item should measure the same unitary construct, in addition to its specific and error components. Unidimensionality is always advisable with tests of ability, but it is more critical for adaptive tests. Thus, a necessary precurser to tailored testing is the demonstration that the item pool is actually unidimensional. Such a demonstration can use existing data for the current ASVAB tests, since it is difficult at present to deal with item response data from tailored tests for purposes of test analyses.

There are several possible ways to obtain evidence of unidimensionality. One way would be to show that the IRT model provided an adequate fit to the item response data. A factor analysis of the item intercorrelations could also give useful evidence. A variety of other methods have been suggested, or under development. Unfortunately each method had drawbacks, so it would be prudent to use more than one method.

Although the theory is based on unidimensional items, empirical results show that the model is suitable when the items have on dominant dimension. Items also related to small secondary dimensions will tend to have smaller *a* values, but will not distort the system.

D1. The fit of the model should be checked. Very desirable.

It might be thought that the fit of the IRT model to the item response data would provide the primary indication of the adequacy of the unidimensional model. After all, the model is unidimensional, so if the model fits the data, the data can therefore be treated as unidimensional, so if the model fits the data, the data can therefore be treated as unidimensional. However, it appears that the fit of the IRT model is not very sensitive to lack of unidimensionality (Jones, 1980). The main mechanism for assessing the fit of the model is to compare empirically-determined item response curves to the curve determined by the item parameters. Items that are multidimensional will tend to have smaller values of *a* (i.e., IRT curves with shallow slopes.) The extent to which these items form meaningful subgroups must be assessed in other ways. This does not mean that the fit of the model is irrelevant to dimensionality. Fit is necessary, but it is not sufficient.

The main reason for comparing the model and empirical item response curves is to check the accuracy of the item parameter estimates and to indicate item idiosyncracies. This recommendation is repeated as a part of the evaluation of methods of obtaining the item parameters. It is relevant here in showing that the model appears to fit.

D2. Highly discriminating items should be selected. Essential.

The main mechanism for insuring unidimensionality is item selection. Tailoring works best when items are highly discriminating, which means that they have high values of a and high correlations with the total test score. Urry (1981) suggests selecting only items with values of a at least 0.8 (on a scale on which ability has mean = 0, standard deviation = 1.) To relate this criterion to more familiar parameters, it can be shown that for c = .25, the biserial correlation of the item and the ability is about .53 when a = .8. This is relatively high for an item-test biserial correlation. Thus, requiring 0.8, or some similar cut-off means that items will have high correlations with ability. Multifaceted items will tend to have lower values of a. Thus selecting items with high values of a will in itself tend to insure unidimensionality. The selection of items with high a-values will also help to make the system efficient, since fewer items will be needed for each person.

There is a danger that a rigid requirement of a=0.8 will force rejection of some good item types. It is well to adjust the requirement to the level that is practically feasible. Also, there may be other reasons such as balanced content, that may indicate including some items with lower a values.

D3. A factor analysis of the interitem tetrachoric correlations should be performed. Very desirable.

In principal, unidimensionality can be examined through a factor analysis of the item intercorrelations. However, determining the item intercorrelations is a problem. Phi coefficients are usually unsatisfactory, because their size depends on the item difficulties, so they tend to yield difficulty factors. Phi coefficients are reasonable when item difficulties are not too disparate, but ASVAB items vary widely in difficulty. A better procedure is to use tetrachoric correlations, although they are not completely satisfactory either. Sometimes the matrix of tetrachorics is not positive definite, thus violating a requirement of factor analysis. In general, large samples of test-takers are needed when using tetrachorics. Also, when the items can be answered correctly by chance, tetrachorics are distorted. Methods of correcting for the distortion have been given by Carroll (1946), Urry (1981), and Samejima (private communication)*. This correction should permit reasonable results from a factor analysis of tetrachorics. However, Reckase (1981) has found that if tetrachorics are overcorrected, results are severely disturbed. By contrast, undercorrection of tetrachorics is relatively safe.

A factor analysis of interitem tetrachorics can be indicative of unidimensionality. Minor departures from unidimensionality will probably not be serious. If there is one prominent factor, and if the secondary factors exhibit either no discernable pattern, or tend to be related to item difficulty, then unidimensionality is supported. By contrast, if there are two or three prominent factors, and if these factors can be rotated (using the oblique or correlated factor model) to show meaningful distinctions, then unidimensionality is challenged.

A variety of similar procedures can be used to assess the factorial structure of binary items. They are all relatively new, and not much experience in their use has accrued. Nevertheless, they are viable alternatives, and are preferable to the above methods. Christofferson (1975) and Muthen (1978) have presented methods for the factor analysis of dichotomous items that are computationally feasible. Another method is proposed by Bartholomew (1980). Bock and Aitken (1981) have suggested a marginal maximum likelihood method, based on multidimensional IRT models that may include guessing terms. The more general approach of covariance structure analysis could also be used. Some of these methods provide chi square tests of goodness of fit, or of the contribution of successive factors added to the model.

D.4. Local independence should be examined. Desirable.

Unidimensionality and the IRT model imply the property of local independence. In the IRT model, ability is the only source of association between items. Thus, for persons with the same ability, the items should be independent. Tests of the local independence hypothesis are being developed by Holland (1981) Levine (Note 1), and Stout (Note 4), but are not ready for practical use. They all rely on indirect assessment, by deducing certain consequences of local independence, and testing the occurrence of such consequences. We are particularly interested in Stout's proposal, and urge trying it when it becomes available.

D5. Subtests should be formed when tests are not unidimensional. Desirable.

Many tests are not perfectly unidimensional. The items tend to cluster by

---

*Samejima's result is as follows. In these formulas, P and p indicate the observed and the modified proportions, respectively, g and $\bar{g}$, or h and $\bar{h}$, denote the correct and incorrect answers to item g or h, and c_g or c_h is the guessing parameter of item g or h, which is unity divided by the number of alternatives.

$$P_{gh}=P_{gh}-[c_h/(1-c_h)]P_{g\bar{h}} - [c_g/(1-c_g)]P_{\bar{g}h} + [c_g c_h/\{(1-c_g)(1-c_h)\}]P_{\bar{g}\bar{h}}$$

$$P_{g\bar{h}} = [1/(1-c_h)]P_{g\bar{h}} - [c_g/\{(1-c_g)(1-c_h)\}]P_{\bar{g}h}$$

$$P_{\bar{g}h} = [1/(1-c_g)]P_{\bar{g}h} - [c_h/\{(1-c_g)(1-c_h)\}]P_{\bar{g}\bar{h}}$$

$$P_{\bar{g}\bar{h}} = [1/\{(1-c_g)(1-c_h)\}]P_{\bar{g}\bar{h}}$$

content area. Achievement tests, and tests of general knowledge are especially prone to such clustering. Such tests can generally be viewed as having a single dominant factor, plus several small group factors, but they can also be viewed as having several highly correlated group factors. When the single dominant factor is sufficiently dominant, or equivalently, when the group factors are sufficiently highly correlated, then the test may be treated as unidimensional. Otherwise, it would be better to treat each cluster, or group factor as a subtest.

The precise criterion for deciding which to do is not easy to specify. A single factor that accounts for 70% of the total common variance is probably strong enough; one that accounts for less than 50% probably signals the use of subtests. If a correlated common factor model is used, we recommend rotating the results to a single general factor plus an orthogonal residual group space, so that these criteria can apply.

Another way of assessing unidimensionality applies to tests in which the items can be sorted into clusters on the basis of item type or item content. In that case, scores can be obtained on the separate subgroups of items. In these cases, for a large heterogeneous sample of test takers, separate scores should be obtained for each subtest, and these subscores intercorrelated. Estimates of the reliability of each subscore should also be made, and the intercorrelation should be corrected for unreliability ("disattenuated"). If the corrected correlations are sufficiently high, (say about .9) the test can be considered unitary.

Finally, when separate tailoring is done, there are two scores which must at some point be combined. Also, if a variable-stopping criterion is used, there are two stopping criteria. Note that we are <u>not</u> proposing to replace one test score by two or more subtest scores when there is some lack of unidimensionality or some separate content areas. No doubt the subtests will be quite highly correlated, and each will be less reliable than a test score should be. Thus the subscores will generally be unsuitable for separate use either in prediction or in counselling and should not be reported separately. Rather the subscores should be combined as suggested. Weights for the combination of subscores into one test score could for example, be chosen so that the resulting score has maximum correlation with the paper-and-pencil version of the test.

We note that comparability with paper and pencil tests may suggest intermixing the items on the subtests. This can be done, in principal, but the equipment must be able to keep track of two or more simultaneous adaptive tests. This is not difficult on general-purpose microcomputers but it is a requirement to keep in mind when obtaining the equipment. This would provide the possibility of multidimensional adaptive testing in the future, that could take advantage of correlations between the underlying abilities while obtaining an estimated score on each.

<u>D6. Tests should be balanced for content and/or item type.</u> Essential. Some tests have heterogeneous content, or use two or more different item

types, or both. The ASVAB includes two tests with heterogeneous content. Auto & Shop Information contains items about autos and items about shop. General science items can be sorted into natural science (physics and chemistry), biology, and health and nutrition. These tests will often be not strictly unidimensional, but may fit the criteria of the previous section. If the test is treated as two or more subtests, then each test taker will necessarily face items of each type, or content. But if the test is administered as a unit, then items should be selected in a balanced way so that as nearly as possible, each person gets the same number of items from each content area. This not only makes the test comparable to the paper-and-pencil version, but it balances the items in cases of any biases for subpopulations. For example, Bock & Mislevy (1981) found that, on the General Sciences Test, to a small extent, males scored relatively higher than females on natural science items, and females scored relatively higher than males on items relating to health and nutrition. Proportional representation of all areas will tend to balance these differences on the test as a whole.

We note that the best way to decide whether or not to balance a particular test for content is by empirical study. Balance may or may not be important. Are there differences between identifiable groups on the different item clusters? Does any such difference imply a relative advantage if content is not balanced on the adaptive test?

When the test is tailored and administered as a unit, balancing the items for content and type requires a hard choice. If the items are selected alternately from the various subgroups, the item subsets will also be nearly balanced for difficulty, but the candidate must keep switching contexts, which would be especially bad with different item types. If, on the other hand, several items are selected from one subset, then several from another, and so on, the candidate need not switch content so often, but content type is somewhat confounded with item difficulty. This is not a severe problem if the test is very nearly unidimensional. We favor this second procedure. Whenever the interaction of difficulty with content could be a problem (because some persons are better on one content, others on another) then separate subtests are preferable. We note again, however, that this is opinion. Empirical evidence is needed to determine the relative advantages and disadvantages of each procedure. There is obviously a limit to how finely the content should be subdivided. Each item is to a large extent specific. There will always be some persons who happen to know more about one item than another. So long as this is either due to the dimension being tested, or is unrelated to that dimension, the specific aspects should tend to average out over a number of items. Careful experimental study of the problem is needed.

One final problem arises in connection with the paragraph comprehension test. In the current ASVAB forms there are several items per paragraph. These items are invariably more highly intercorrelated than are items from different paragraphs. This violates the principal of local independence of items, which is central to IRT, and hence to CAT. Thus in tailored testing, ideal items would have only one question per paragraph. Efficiency would then require short paragraphs. We note elsewhere that the capacity of the display screen limits the length of the paragraph

comprehension questions. However, shorter items, with one question per item may measure a somewhat different skill. Empirical evidence on the comparability of item types is important. If empirical research shows that there is noticeable difference, the P&P version of the ASVAB should be changed to match the CAT version.

Finally, if multiple-item paragraphs are used in the CAT version, some form of multiple response model should be used that does not assume conditional independence among responses to the same paragraphs (e.g. Samejima, 1969). In this case, tailoring will have to operate with respect to paragraphs rather than items, but the greater precision of these types of items may allow the number of paragraphs presented to be less than the number of items presented in those tests where all items are independent.

Reliability and Measurement Error

In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance in the population of persons from which the examinees are assumed to be randomly sampled. This quantity can be expressed as the intraclass correlation,

$$\rho = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2},$$

where σ_x^2 is the observed score variance and σ_e^2 is the measurement error variance. The correlation, ρ, is estimated directly by the customary indices of reliability, such as parallel-form or test-retest reliability, and indirectly, by split-half reliability; Cronbach's alpha provides a lower bound to .

By a simple algebraic manipulation, the measurement error variance can be expressed as

$$\sigma_e^2 = \sigma_x^2(1-\rho);$$

the standard error of measurement is then

$$\sigma_e = \sigma_x\sqrt{1-\rho}.$$

Although the reliability is a convenient unitless number between 0 and 1, the standard error of measurement is more useful in score interpretation.

The formulas above are used almost universally in test practice, but they make use of the generally false assumption that the error variance is the same for all scores, rather than being dependent on ability. It is widely recognized that the measurement error is not constant for conventional tests, being larger at the extremes of the ability distribution and smaller near the mean. Since the classical formulation above uses a single average value for the error variance, σ_e^2, the conventional reliability coefficient is at best a crude description of the true state of affairs.

As Samejima (1977a) has pointed out, this definition of reliability has little relevance for measurement based on item response theory, where

the error variance is expressed as a function of ability. In item response theory, the estimate of error variance is expressed as the variance of estimated ability, $\hat{\theta}$, for a fixed value of ability, θ. The estimate of error variance will depend on the method used to estimate θ. Using classical statistical theory, with maximum likelihood estimation of θ, the variance of measurement error is given by the reciprocal of the *informaton formation*, as noted in the introduction. In Bayesian theory, the error variance is also readily computed.

In adaptive testing, the measurement error variance depends on the stopping rule. One stopping rule is based directly on the error variance or its reciprocal, the information function: all examinees are tested to the same value of the error variance, or information, over as wide a range of ability as practical. In that case, the estimated standard error of measurement is constant. If the item pool is not large enough to support a uniform information criterion, or if some other stopping rule is used, such as a fixed number of items, then the error variance will still depend on the ability level, θ.

One further practical problem in the use of IRT is the score scale. The natural scale for IRT theory is the ability scale, θ. This scale is non-linearly related to the conventional score scale, which is used on the paper-and-pencil ASVAB. Since, at least at present, it may desirable to transform the CAT scores to expected number-right scores, the standard error of measurement will again be a function of the score value. The expected-number-right score is a strong monotonic function of θ, and is given by

$$x = \sum_{j=1}^{n} P_j(\theta)$$

then

$$\sigma_{\hat{x}|x} = \frac{dx}{d\theta} \sigma_{\hat{\theta}|\theta}$$

where

$$\frac{dx}{d\theta} = \sum_{j=1}^{n} \frac{dP_j(\theta)}{d\theta} .$$

This discussion leads to the following recommendation.

E1. The standard error of measurement of each test score should be reported as a function of the test score, in the metric of the reported score. Either a graphical or tabular report form or both, can be used. Essential.

E2. The standard error of measurement of each test should also be reported in the ability metric, as a function of the test score unless the *standard error is constant*. Desirable.

This recommendation is for the convenience of further psychometric analysis of the test. This information will be useful only in connection with the actual item parameters, and should be kept separate from the report in E.1.

Reliability. Custom suggests that a unitless index of reliability also be provided, although such an index is somewhat contrived. Many psychometricians feel that devising a reliability coefficient for an adaptive test is inappropriate and misguided. Nevertheless, if one is determined to have a reliability coefficient in item response theory, there are two possibilities. One approach is to define a conditional reliability, i.e.,

$$\sigma(\theta) = \frac{\sigma_x^2 - \sigma_e^2(\theta)}{\sigma_x^2}$$

which could be graphed or specified at selected values of θ. This would be the reliability if everyone were measured with the same precision as those persons with ability θ. This function is somewhat like the information function except that it has the convenient property of being unitless.

The other possibility is to define an average or marginal measurement error, in a population with ability distribution $g(\theta)$,

$$\sigma_{em}^2 = \frac{\int_{-\infty}^{+\infty} \sigma_e^2(\theta) g(\theta) d\theta}{\int_{-\infty}^{\infty} g(\theta) d\theta}$$

where, if ability is normally distributed, these integers can be evaluated by Gaussian quadrative, as discussed by Bock & Lieberman (1970). Then marginal reliability can be defined as

$$\rho = \frac{\sigma_x^2 - \sigma_{em}^2}{\sigma_x^2}.$$

E3. Reliability should be reported by the marginal reliability index. If testing is to a fixed error criterion, this is equivalent to classical reliability. Very desirable.

E4. Conditional reliabilities should be reported at selected points on the ability scale. Very desirable.

E5. The precision of P&P and CAT versions of the ASVAB tests should be compared. Essential.

For comparative purposes it will be necessary to show the precision of the current P&P (paper and pencil) versions of the ASVAB. We recommend finding item parameters for one current form of each test to obtain the test information function and thus the measurement error variance. It is

recognized that the CAT will be more efficient because of its adaptive character and also because the stringent criteria for item selection will result in more discriminating items on the adaptive test. Still the important practical question is the overall gain. Such a demonstration will not be difficult. IRT item parameters have been obtained for at least one form of the ASVAB, both by Ree (private communication) and by Bock & Mislevy (1981). Of course the parameters must be obtained by the same statistical procedure that is used for the CAT items, so additional work may still be needed.

Eventually it will be desirable to compare the precision of each CAT with the precision of a hypothetical P&P test formed from the CAT item pool, picking items that matched the actual P&P test in difficulty but matched the CAT pool in discrimination. Such a comparison would indicate how much of CAT's efficiency is due to better items and how much to tailoring.

Empirical reliability.

The above definitions refer to error due to sampling of items from an indefinitely large pool. They do not include variability due to short-run random variation of the trait being measured or to situational variance in the testing conditions. These sources of error can only be assessed empirically. It would be desirable to estimate the extent of this type of variation by readministering the adaptive test on successive days or weeks, with the condition that items presented to the same subject are sampled without replacement. Pearson product-moment correlations of the paired measurements would serve to estimate the empirical reliability. Because the items are not repeated, the reliability determined in this way is equivalent to classical alternate-form reliability. Because of the way the items are selected, this reliability might be called stratified, randomly-parallel form reliability. Because the scores are obtained on different days, test fatigue would be avoided; because the days are close in time, this reliability coefficient would indicate the short-term stability of the scores.

E.5. Alternate-form reliability should be determined empirically for each test in the battery. Essential.

It should be noted that most decisions are based not on individual test scores but on various composites. Initial entry, for example, is based on the AFQT composite. Although there are too many composites to enable calculating reliabilities for each composite, reliabilities can be computed for the most widely used composites, and the data should be available for computing reliabilities of arbitrary composites. This requires the reliabilities of the individual tests, and the intercorrelation matrix of the tests.

E6. Reliability of widely used composite scores should be reported. Highly desirable.

E7. Test intercorrelations should be reported. Highly desirable.

Validity and differential prediction.

Before switching from paper and pencil (P&P) versions of the ASVAB to a computerized adaptive test (CAT), it is important to have evidence that the CAT is at least as valid as the P&P version of the battery. During a transition stage when P&P and CAT are both in operational use, it would also be important to have evidence that scores on the two test forms have the same predictive meaning. Included in the latter category would be investigations of possible differences in CAT prediction equations for key subpopulations (e.g., differential prediction as a function of race or sex).

Comparisons of variance-covariance matrices and covariance structures are needed. A comparison of interrelationships among the tests for the two forms would provide an initial check for possible differences in the validities of the CAT and P&P tests. Correlations among the tests may be altered due to differences in the precision of measurement at different ability levels, with the CAT version expected to yield better measurement at the extremes. When based on different samples, the correlations would also be expected to vary as a function of group heterogeneity. For these reasons, comparisons of variance-covariance matrices and of covariance structures will be more informative than comparisons of correlation matrices.

The predictive validity of the P&P version of the ASVAB has been well-documented (Department of Defense, 1980; Fischl et al, 1978.) If the CAT versions of the tests are highly related to their respective P&P counterparts, and if the covariance structures are similar, similar predictive validity can be inferred. There are various opinions about potential differences in validity for adaptive tests in general. The CAT tests may possibly be more nearly unidimensional than the conventional P&P tests, but the purity may be seen as clarity, implying improved validity, or as sterility, implying poorer validity. Kingsbury & Weiss (1981) and Sympson & Weiss (Note 6) claimed to be optimists, but in fact found very little difference between the validity of the two modes.

V1. The similarity of variance-covariance matrices should be assessed. Essential.

The most straightforward comparison is a test of the hypothesis of the variance-covariance matrices for the CAT and P&P versions, $\Sigma_c = \Sigma_p$. Procedures for testing this hypothesis are well known (e.g., Box, 1950). The main requirements for the purposes of the CAT vs. P&P comparison is that (1) the two versions of the tests are administered to random samples from the same population, and (2) that the number of examinees taking each version is relatively large (say, 200 or more).

Assuming that the CAT will be experimental, it will be subject only to incidental selection. Comparable P&P data would then require a special administration (retesting) with another P&P form, so that P&P and CAT scores are both subject to incidental selection, and so that both scores are obtained in retesting, with as nearly similar motivational conditions as can be managed, and with order of testing counterbalanced. Other

experimental designs are, of course, possible.

Although the comparison of variance-covariance matrices is important it will not provide a direct indication of the source of any differences that are found. For that purpose, the comparison of covariance structures outlined in the following section should be more useful.

V.2. The covariance-structures of the two versions should be compared. Very desirable.

It is recommended that the factor structures of the CAT and P&P versions be compared. Using subscripts p and c to designate the P&P and CAT results respectively, the general form of the two n x n variance-covariance matrices would be:

$$\Sigma_c = \Lambda_c \Phi_c \Lambda_c' + \Psi_c ,$$

and

$$\Sigma_p = \Lambda_p \Phi_p \Lambda_p' = \bar{\Psi}_p .$$

After some exploratory analyses of results for a large sample of P&P results, a hypothesized pattern of zeros and free parameters in p would be determined for m n factors.

In the initial data collection, it may be necessary to limit the computerized testing to the seven of the ten ASVAB tests that do not involve graphics. Those seven tests might be hypothesized to have a pattern as shown below with X's indicating free parameters and zeros fixed parameters.

Initial comparison of factor patterns and structure will be limited to seven tests. When graphical capabilities are added, the analyses illustrated above should be repeated with the complete battery of ten tests. For the ten-test battery, we expect the following 4-factor pattern. Here we separate large loadings X, medium loadings x, and zero loadings, 0. In a confirmatory analysis, all x's would be free parameters.

Test	Factor 1	2	3	4
General Sciences	0	X	0	x
Arithmetic Reasoning	X	0	0	0
Word Knowledge	0	X	0	0
Paragraph Comprehension	0	X	0	0
Numerical Operations	0	0	X	0
Coding Speed	0	0	X	0
Auto & Shop Information	0	0	0	X
Mathematics Knowledge	X	0	0	0
Mechanical Comprehension	x	0	0	X
Electronic Information	0	x	0	X

1. Equal factor loadings: The first constraint to be imposed is that $\Lambda_c = \Lambda_p$. All variances and covariances in Φ_p and Φ_c would be free and not constrained to be equal. The matrices would also be unconstrained diagonal matrices.

2. More constrained models: Additional constraints that $\Psi_c = \Psi_p$ and/or that $\Phi_c = \Phi_p$ could also be added. It seems likely, however, that different matrices would be required.

It is assumed that Form 8, 9 or 10 will be used for the P&P tests. Comparisons will also be needed between CAT and Form 11, 12, or 13 before the CAT is made operational. This is important because if the CAT becomes operational it will be used along with Forms 11, 12, and 13. Thus, the above comparisons should be repeated using Form 11, 12 or 13 for the P&P version. Alternatively, the comparisons of all three versions could be made simultaneously by administering Form 8, 9, or 10 to one-third of the sample, Form 11, 12, or 13 to one-third of the sample and the CAT to the remaining third.

As in the simple comparison of the covariance matrices, it is important that the results be based on sizeable random samples from the same population. Several available computer programs are capable of performing the above analyses. One of the better known programs is LISREL V (Joreskog & Sorbom, 1978). In addition to providing chi-square tests of the hypotheses suggested above, standard errors of the parameter estimates and residual differences between the sample variances and covariances and those estimated by the model may be obtained. Both the standard errors and the residuals should be reported. They will be useful for purposes of

judging the practical importance of any statistically significant differences that are obtained.

V3. The CAT battery should be validated using external criterion measures. For comparison, the P&P battery should be validated using the same criterion measures. Essential.

The analyses of the covariance structures as outlined above will provide a good test of the extent to which the CAT and P&P versions are measuring the same abilities. It will nonetheless be important to compare directly the prediction equations of the CAT version with those developed for the P&P version with a few important criterion measures. If nothing else, it would be useful for purposes of satisfying skeptics. Such comparisons will be essential, however, if the two versions of the tests are found to have different covariance structures. The latter outcome also seems quite likely since the CAT will have nearly equal precision across the range of test scores, whereas the P&P test is much less accurate at the extremes relative to the middle of the ability distribution.

For several reasons, comparisons of CAT and P&P correlations with criterion measures may not be satisfactory. As was just indicated the precision of relative efficiency of the two versions is apt to differ as a function of ability level. Also, the samples for which criterion data could be obtained will generally be subject to explicit selection on the P&P tests but only incidental selection on CAT. Thus, the P&P correlations would be expected to be affected more by selection effects than the CAT correlations would be.

The comparisons of primary interest can all be classified under the heading of differential prediction. Comparisons of regression systems, including error variances, slopes and intercepts are all relevant. Comparisons of two types of regression equations should be made: (1) regression of a criterion measure on subtest scores and (2) regression of a criterion measure on test composite scores. To the extent that it is feasible, it would also be desirable to compare the conditional variances on the criterion measure as a function of test score. It might be expected that the conditional variances would be smaller for extreme CAT scores than for the corresponding scores on the P&P, whereas the conditional variances would be more nearly equal in the middle. The possibility of nonlinearity would also be worth investigating to the extent that this is feasible.

Although differences in error variances would be a concern, the more serious concern would be with differences in slopes and/or intercepts. The latter types of differences would imply that systematic errors of prediction would result from using CAT scores in place of P&P scores. For schools that have minimum entry requirements, the predicted criterion scores for test scores near the minimum deserve special attention. If the predicted value on the criterion is significantly higher (lower) for the CAT than for the P&P, then individuals who take the CAT would be given an unfair disadvantage (advantage). The Johnson-Neyman (1936) procedure could be used to determine if the minimum score fell in a region of significant differences. If there are significant differences in predicted scores associated with test scores at the cutoff, then different entry

requirements would be needed for the two test versions. Such a finding would also suggest that a more comprehensive series of differential prediction studies would need to be undertaken to determine the generalizability of differences in prediction for other training areas.

The schools and criterion measures to be used will need to be determined on the basis of feasibility and importance of selection for the various specialty schools. They should have some variety (e.g., auto mechanic, clerk-typist, electronics and infantry). It is recommended that differential prediction studies be conducted for at least three schools, and preferably many more. For each school there should be a minimum of 100 persons with CAT scores and another 100 or more with P&P scores. The P&P scores should be obtained from a special administration rather than from the files in order to avoid differences due to motivational differences for the special administration in comparison to regular administration at time of entry into the service. Final course grades, performance tests and attrition might all serve as criteria. For selected jobs in the Skills Qualification Test results might also serve as criteria.

The use of only 100 cases in each group will be adequate to detect gross differences in prediction equations and will be satisfactory as a first step. If at all possible, 200 cases would be much better. Still, small differences in regressions cannot be detected with fewer than 500 cases in each group, and subtle differences need even more cases, or a different approach to aggregation. We recommend that attempts be made to gather enough data for such comparisons, although we realize that this could not happen before widespread use of the CAT battery.

V4. The extent of prediction bias should be assessed for important subpopulations. Desirable.

It would be desirable to compare prediction systems based on samples from important subpopulations. Of special interest is the possibility that a prediction system based on results for men yields biased predictions for women or that one based on majority group results yields biased predictions for Blacks or Hispanic persons. The major obstacle to investigating the possibility that the CAT leads to biased predictions for members of particular subpopulations will be sample size. For useful comparisons it is desirable that CAT results and criterion results be available for approximately 100 or more members of each subpopulation. If it is feasible to obtain samples of this size for a particular school, then standard tests for the homogeneity of error variances, slopes and intercepts should be conducted. If significant differences are obtained, then the direction and amount of bias would need to be examined as a function of scores on the CAT, and requires even more cases. Bias in prediction near the minimum score for entry into a school would be of special concern.

Item Parameters - Estimation

The parameters of the item response curve for each item in the test pool play a central role in adaptive testing. The choice of items to present to each person, and the score derived for each person depend critically on the item parameters. Without good items and good estimates of the IRT parameters, useful ability estimates will not be obtained, regardless of the quality of the other components. The item parameter estimates are used for item selection, ability estimation, and to compute test information. If the parameter estimates are poor, none of the other procedures can give meaningful results. Therefore it is of utmost importance that the calibration be done properly and that evidence be presented to show the quality of the results.

The following sections will make recommendations concerning how estimation, linking, equating, and item pool production should be done, based on the best current information and judgment. New research in the area may require changes.

Item calibration. The term item calibration is used here to mean the estimation of IRT item parameters for each item in the item pool for a test. These parameter estimates are usually obtained from one of several calibration programs that are available. Following the background discussion, it will be assumed that a is the discrimination or slope parameter, b is the difficulty, or threshold parameter, and c is the lower asymptote, or pseudo-guessing parameter. These parameters will be assumed to be in normal ogive form. That is, if a logistic model is used, the constant D = 1.7 is included in the model, as in the presentation given above in the section labelled "Background."

The primary requirement in determining item parameters is having enough cases to yield stable estimates. Although the sample size requirements for the various calibration programs vary, the current literature (Lord, 1968; Reckase, 1978; Ree, 1979, 1981) seems to indicate that at least 1,000 cases are necessary for stable calibration. This a firm lower limit. A larger sample is desirable. Our general recommendation follows.

IE1. The sample for item calibration should be of adequate size, currently at least 1000 cases. Essential.

As a corollary, any new procedure for item calibration is likely to need the same sample size. However, the requirement of 1000 cases is the result of empirical test. Thus, when considering a new procedure, the sample size requirements must be reevaluated using both simulation and live data studies. For the simulation studies, samples of item response vectors should be generated using the model selected as a basis for the CAT system for the test length to be used in item calibration and using realistic assumptions about error. Several different sample sizes should be produced so the effect on calibration can be determined. These samples should then be used to determine the item parameters of the simulated items. These estimated parameters can then be compared to the item parameters used to generate the data to determine the adequacy of the sample size. Both squared deviation and absolute deviation statistics have been used for the

comparison in the past. Another check that has not frequently been used in the literature, but that we advocate, is to compare empirical and theoretical item response curves. In estimating parameters, one also estimates ability values for the persons, which then permits determining the empirical curves for comparison.

The live data studies can be performed by calibrating a test on a large sample that is well beyond the sample size expected to be required for accurate calibration, and use those results as a basis for evaluating the quality of smaller sample calibrations. As with the simulation procedure, a squared deviation or absolute deviation statistic can be used to judge the similarity of the parameter estimates from the small sample and large sample calibrations. The goal is to determine the point where an increase in sample size does not produce any meaningful increase in similarity. What is a meaningful increase is still a subjective judgment.

Both simulation and live data studies should be run to evaluate calibration procedures because of the basic inadequacies inherent in each type of study. Simulated data never accurately represent the many extraneous sources of variation present in real data. Therefore, simulations tend to give a better result that can be obtained from real test data. By contrast, studies using real data have the problem of not knowing the "true" parameters, so they lack a good criterion for accurate calibration. Results from extremely large samples do not provide the criteria, because they may be biased if a poor calibration procedure is used. By using both simulated and real data, the weaknesses of each type of study can be taken into consideration, resulting in an estimate of the required sample size that can be accepted with greater confidence.

Merely having a large sample of examinees is not sufficient to ensure that calibration results will be accurate. If the ability of the sample is such that most examinees have a high probability of responding to the items - that is, the test is too easy--it will not be possible to estimate two critical parameters of the item response curve. These parameters are the ones dealing with the lower asymptote (guessing level, c) and the slope at the point of inflection of the curve (discrimination, a). In order to estimate these parameters, the sample must have sufficient numbers of cases at the middle and bottom end of the ability range measured by the items. Thus, a large sample that is positively skewed is more desirable than one that is negatively skewed. If necessary, the tryout sample should be specifically chosen to have sufficient cases in the middle and lower ability ranges.

IE2. The calibration sample should be selected so that a sufficient number of cases are available in the range of ability needed to estimate the lower asymptote and the point of inflection of the IRC. Essential.

The statistical properties of the item calibration procedure should be carefully evaluated. Since the selection of items and the estimtion of ability are both totally dependent on the accuracy of the item parameter estimates, it is of critical importance that the estimates be shown to be good approximations of the "true" parameter. From a statistical point of view, there are several criteria for what is considered a good estimate.

The two criteria considered of importance here are measures of consistency and statistical unbiasedness. A consistent estimate is one for which the expected values of the estimates will approach the true value as the sample size increases, and its variance will approach zero. This criterion ensures that large sample estimates are good estimates. An unbiased estimate is one for which, at any sample size, the expected value of the estimate equals the true estimate. A biased estimate can be consistent, if the bias gets smaller as the sample size increases, and tends to zero as the sample size increases without bound. There is no question that the estimates should be consistent, but there is some argument about bias. Further, there is no theoretical proof at present that any of the methods yields consistent estimators. Establishing consistency is at present an empirical problem, so we use the term empirical consistency.

Bias may be less important. All Bayesian estimations are biased Such estimators may have smaller mean squared error than other methods, in which case their use is justified. But, when biased estimation is used, the extent of the expected bias should be known.

These considerations lead to the following recommendations.

IE3. The procedure for estimating item parameters should be shown to be empirically consistent. Essential.

IE4. The procedure for estimating item parameters should be shown empirically to be unbiassed, or the extent and nature of the bias should be specified. Essential.

Bias is a problem mainly in putting together estimates obtained from different data sets. Such combinations of estimates is required in adaptive testing, because a large item pool must be calibrated for each test. If equivalent samples of the same size are used in calibrating different sets of items, the calibrations can be linked in a straightforward way. But if the samples vary in sample size or in the shape of the ability distribution, biases may differ, introducing extra error in the linking process. The bias can also be troublesome if items are recalibrated in an operational setting, which we recommend below, for reasons presented there. Here the issue is that any procedure for recalibrating the items will have to recognize the inherent bias in item parameters of the item calibration procedure is biased.

We note that the issue of estimation bias is critical because some of the prominent procedures for item calibration including one due to Urry (1981), uses a Bayesian framework, which is inherently biased. This statistical bias is not seen, by Bayesians, as a bad thing but as a conservative thing, in the same way that ordinary least-squares regression yields conservatively biased predictions. In essence, the estimates are biassed toward a prior distribution of ability, which is commonly specified as normal. Although Urry's procedure does not include prior distributions for item parameters, the net result is that the b parameter estimates tend to be regressed toward the mean ability. The procedure of Swaminathan (Note 5) and Reiser (Note 3) does include prior distributions on c and a. So long as bias can be measured explicitly in all uses of the parameters, the bias can be tolerated, but it does complicate the system.

Results from simulation studies can be used to determine if the parameter estimation procedure yields empirically consistent and unbiased estimates. Response data can be generated for a sample similar in size to that available for live testing applications using specified item parameters and a known ability distribution. This data can then be calibrated using the estimation procedure of interest and the parameter estimates compared to the known, true parameters. This can most easily be done by plotting one set against the other. If the resulting plot tends to follow a 45 degree line, the estimates are unbiased. If the plotted points cluster more closely around the line with increased sample size, the estimates may be called empirically consistent. An alternative analysis is to compute average squared deviation or absolute deviation statistics between the true and estimated parameters to indicate their similarity.

It is recognized that the guessing level parameter, c, is not easily estimated for easy items. There will be a need for items that are so easy that even the lowest-scoring persons will have a moderate probability of correctly answering the item. For those items, estimate c is very difficult. Recent work by Swaminathan (Note 5) and Reiser (Note 3) on Bayes-constrained estimation of the parameters has improved prospects for stable estimation of the c parameters.

Simulation data alone cannot demonstrate the effectiveness of an item calibration procedure. No matter how conscientiously produced, simulation data does not have the same richness of variation as the responses of individuals to test items. Therefore, it is important that the calibration be shown to yield satisfactory results on real data as well as simulation data. The procedure used to determine the quality of item calibration is the comparison of empirical item response curves (IRC's) with the IRC's based on the item parameter estimates. Empirical IRC's can be obtained by dividing the ability scale into several intervals and determining the proportion correct for each interval from the item data. A considerable number of intervals should be used (we suggest 15-20) so that the variation in ability within an interval is small enough to be ignored. Both the empirical and estimated IRC's can then be plotted on the same axes for comparison. A quantitative evaluation of these curves can be obtained by using the chi-square statistic suggested by Yen (1981). Strictly speaking, this statistic should be used only when the abilities are estimated from other items, but it does give a means of judging the relative fit of the estimated IRC's to the actual item data. Levine (Note 1) has also proposed a method of assessing the fit of IRC's.

IE5. The IRC's defined by the estimated item parameters should fit the observed data. Essential.

It will probably be necessary to do the initial item calibration with data obtained from P&P administrations. It is possible that the characteristics of items are different in the P&P and the CAT formats. A study is suggested below, in the discussion of human factors, to examine this issue. If there should be an effect of mode of presentation, this effect will have to be taken into consideration when equating the CAT scale with the current P&P ASVAB scale.

When CAT becomes operational, or as early as may be, a check should be made of the item parameters in the operational context. This may involve a re-estimation of item parameters, and at least should involve a comparison of the IRC curve specified by the item parameters and the empirical IRC curve. To do this involves inserting each item into the series for a group of examinees, since if we rely on the normally accumulated responses for the item, the upper and lower portions of the curve will be poorly estimated.

IE 6. The operational CAT system must be able to include a specified item in a test sequence without scoring it, on a flexible predetermined schedule. Essential.

When this has been done for all or most of the items in the pool, the data should be examined to determine if some adjustment in item parameters is necessary. Here the question of test score calibration is paramount.

IE7. Items that include diagrams should be recalibrated either on prototype equipment or on the operational equipment. Essential.

There will be enough uncertainty with the tests containing diagrams that such items should be recalibrated on the actual equipment. Some tests use items with diagrams, as discussed below under human factors. Difficulty of the item may be altered by the legibility of the diagrams.

IE8. As soon as possible a study must be done to compare the difficulty parameters of items given in the standard paper and pencil mode with the same items given by the computer. Essential and urgent.

One overall effect of computer presentation may be to change the difficulty of items on the power tests. The effect on the speeded tests is certain, as noted elsewhere. If the effect on the power tests is significant, it will have implications for the plan to calibrate the item pools by P&P methods. If the effect is constant across items, it will not be noticed, since the item calibration is relative to an ability distribution with a specified mean and variance. A constant effect would, however, cause trouble in equating with previous tests, so the actual size of the effect must be determined. But if some items are affected more than others, item parameters determined from the P&P mode are open to question. We do not expect a differential effect, except possibly for items with diagrams. But an empirical determination of the presence or absence of a differential effect is necessary.

The experiment should be done first with power tests that do not include diagrams, using experimental or prototype equipment. When appropriate equipment becomes available, the tests with diagrams should be examined.

The experiment should compare the two modes of test presentation, in the context of a standard test, with the computer not in an adaptive mode, because data obtained from administering an adaptive test cannot readily be used to estimate item parameters. In an adaptive test, each item is administered to a different set of persons, usually whose probability of giving a correct response is neither very small nor very large. Comparability of results in the proposed experiment requires that the people who attempt each item in the two conditions have the same ability distributions.

Note that any single item can be calibrated in an adaptive setting by administering it to all persons, independently of their ability, randomly inserting the item into the sequence of administered items. But this procedure would require large numbers of examinees in the present study, since a different group would be needed for each item.

The experiment should balance order of presentation, for the special battery of power tests, being examined. All tests of the speed battery should be given together in one mode, and then in the other mode. (It is conceivable that the adaptive test would in itself sufficiently change the difficulty, but this seems most unlikely. The change due to mode of presentation is also unlikely, but is at least conceivable.) Many experiment designs are possible. In one, which seems to be the simplest, each test in the battery would be prepared in both P&P mode and computer presentation mode. Two groups of subjects would be randomly assigned (this is vitally important) to one of two groups. Group I would take the battery in P&P mode, Group II would take the battery in computer presentation. An analysis of variance would be run on the parameters themselves, using the b values directly, but using log of a and logit of c. These transformations will make the data appropriate for the linear analysis of variance model. Notice that sample size need not be 1000, as in ordinary parameter estimation, because there is no intent of using the parameters for individual items. The issue is whether there are main effects and interactions for the set of items. Probably samples of 200 in each group would suffice.

The main question is whether there is a main effect on item difficulty and whether item difficulty interacts with test mode. The intercorrelation of the two test modes is a secondary aspect of this question. It would be best to fit IRC's to these data, but the main questions can be answered by standard item analysis.

Item parameters - Linking.

Linking is the process of putting the results of separate calibrations on the same scale. Unless large numbers of items can be administered to many individuals, linking is necessary for the formation of large item pools. The usual procedure in forming large item pools is to administer many short tests to many different groups of individuals. The results of the separate calibrations of each test are then linked together to form one large set of calibration data. Since the parameter estimates actually used in the CAT procedure are those determined from the linking, the quality of these estimates is critical. Even good calibration results can be ruined by poor linking.

In a recent study of linking procedures (Vale et al, 1981) four different types of linking designs were considered. Two of the designs depended on sampling. In the equivalent-groups procedure different subsets of items are given to different random samples of the population of test-takers. Each set of items is calibrated separately, and the results rescaled so that the mean and standard deviation of the ability scores of the two groups are equated, on the assumption that the groups are equivalent.

In the equivalent-tests method, subsets are determined by a random process, and are given to different groups. It is assumed that the process results in parallel tests. We doubt the wisdom of such an assumption, since small samples of items are involved. Since Vale et al found this method inferior, it is not recommended.

Vale et al also consider what they call the anchor-group method in which one group of persons takes all the items. Since the point of linking is to avoid such a requirement, this method is not recommended.

The other viable methods involve overlapping sets of items. The most common such design is the anchor-test method in which one subset of items is taken by all persons, and provides the base for linking the remaining items. This design is sound but is especially pertinent to equating successive forms of a test in a testing program like the College Board series. An extension of this design has each test sharing some items with some other tests. (See McKinley & Reckase, 1981b.) The optimal design for use in this method is a balanced incomplete block design where the groups of individuals define the blocks and the items are the treatments. Each test would be calibrated separately for each group and the parameter estimate would be used as the dependent variable in an analysis of variance to determine the transformation (treatment effect) required to place them all on the same scale. If the equivalent groups method is used with parameter estimates from a three-parameter model, the b-values can be used as is, but the log of the a-values should be used and the logit of the c-values, as suggested above. As an alternative, a program that accepts a not-reached code (such as LOGIST or BILOG) can be used to estimate all items simultaneously.

IL1. When using a common item procedure to link calibration together, the parameters must be shown to be on the same scale. Essential.

The procedure recommended to show the quality of the linking procedure is to form a circular chain of linked tests with the first test eventually linked to the last test. That is, Test 1 should be linked to Test 2, Test 2 to Test 3, Test 3 to Test 4, and Test N-1 to Test N where Test N is the same as Test 1. Thus the initial parameter estimates for Test 1 can be compared to the linked parameter estimates for Test 1 (Test N) to determine their similarity. This procedure can be performed using data from the balanced incomplete block design described above.

IL2. The linking procedure used should be fully described. Essential.

IL3. The similarity of initial and linked estimates should be presented. Essential.

Correlations are inappropriate for this measure of similarity. The parameters could be on quite different scales and still give high correlations. Some type of deviation statistic such as the average squared or absolute deviation would be much more appropriate.

IL4. When using an equivalent group procedure to get the parameter estimates on the same scale, the groups used must be shown to be equivalent. Essential.

The equivalent group procedure is totally dependent on the similarity of each of the groups used for calibration. The sampling plan for obtaining equivalent groups is critical.

IL5. The methods for sampling the individuals for the groups should be described in detail. Essential.

IL6. Descriptive statistics showing the equivalence of the groups should be reported. Essential.

The means and standard deviations alone are not enough to show the similarity of groups for IRT linking. The distributions of scores must be shown to be similar. This can be done by reporting coefficients of skewness and kurtosis or by graphically comparing distributions.

Item Pool Characteristics.

Item pool characteristics include the placement and quality of items along the ability scale. Regardless of the quality of the calibration and the linking, if good items are not present in the region of the ability scale of interest, good ability estimates will not be obtained. This fact is reflected in the size of the standard error of the ability estimates or in the number of items needed in order to achieve a specified error criterion.

IP1. The distribution of the a-parameter estimates and descriptive statistics for the estimates should be presented. Very desirable.

IP2. The distribution of the b-parameter estimates and descriptive statistics for the estimates should be presented. Very desirable.

Special mention should be made of any gaps in the item pool.

IP3. The distribution of the c-parameter estimates and descriptive statistics for the estimates should be presented. Very desirable.

IP4. The information function for the total item pool should be presented. Very desirable.

The information function will show where the item pool has adequate numbers of items and where few or low quality items are present. It will also indicate the range of ability that can be measured by the pool.

IP5. The anticipated ability distribution should be plotted on the same scale with the information function. Essential.

Even if the item pool is of good quality, it will not result in good measurement unless it matches the ability of the population of interest. For example, if a large number of difficult items are administered to a low ability group, poor measurement will result regardless of the quality of the items. One way to easily check if the items match the ability of the groups is to plot the ability distribution of the examinee population on the same graph with the item pool information function. The two plots should overlap for the majority of their range. If no existing distribution exists, a normal distribution with mean and standard deviation estimated from past testing can be used.

Item Selection and Test Scoring

Several different methods might be used to select successive items in adaptive testing. The up-and-down method of stochastic approximation (see Lord, 1970) adjusts the difficulty of the next item either up or down a fixed amount, called the step size. The Robbins-Munro procedure (see Lord, 1971, and Sampson, 1976) provides a method for decreasing the step size as the testing progresses. Neither of these methods is advocated now, because more powerful procedures are computationally practical.

Three distinct methods are presently available for computing provisional estimates of ability and selecting items for sequential testing - (1) the Bayes updating method proposed by Owen (1969, 1975), (2) the maximum information method discussed by Lord (1977) and Samejima (1977a,b), and (3) a finite Bayes method recently proposed by Bock and Aitkin (1981). The principles, advantages, and disadvantages of each are discussed in this section.

1. Bayes updating. Although informal trials of adaptive sequential item testing had been carried out earlier (Linn, Rock & Cleary, 1969), the first statistically motivated proposal was that of Owen (1969, 1975). He gave a Bayes updating rule based on the posterior mean and variance, given the subject's response to one item. On the assumption that the posterior distribution is approximately normal (strictly speaking it cannot be exactly normal even when the prior is normal), Owen's result can be applied recursively to estimate the mean and variance of the posterior distribution after any number of successive item responses. The mean is then the estimate of the subject's latent ability and the variance, the estimated measurement error.

The item selection rule is to choose the item that will most reduce the posterior variance. That item proves to be the one with the highest discriminating power among those in the neighborhood in the prior mean in difficulty. The process is repeated until the stopping criterion is reached.

Owen (1975) proved that this rule almost certainly converges to the value of the trait, and Wood (1971) demonstrated its properties in application to real and simulated subjects. Using a 1000-item pool of vocabulary items, Wood successfully estimated vocabulary knowledge of high, medium and low ability 4th, 5th, and 6th graders with uniformly good precision with about 25 items in most cases. He found, in fact, that better precision than with conventional tests was often attained with twenty items, the gains in precision being small after that point. McBride (1977) also studied Owen's strategy.

The equations of Owen's method are relatively simple because of the use of a normal prior distribution and normal item characteristic curves, and by the simplifying assumption that the posterior distribution of ability is also normal.

Recently, Bock and Aitkin (1981) have shown that a straightforward

Bayes method can be used, without the simplifying assumption, because the necessary, and apparently complicated, calculations are, in fact, quite simple to do by numerical methods of integration.

2. Maximum information. The result from item response theory most crucial to adaptive testing is the provision for "item-invariant" estimation of ability. Estimates on a common scale can be computed for different examinees for different subsets of items from a calibrated item pool. Thus, items that are optimally informative can be selected for each examinee without affecting the comparisons between examinees. In the class of item-invariant estimators, the maximum likelihood estimator proposed by Birnbaum (1968) has been most intensely investigated for use in adaptive testing. Lord (1977b), Samejima (1977a), and others have used Monte Carlo methods to examine the properties of the maximum likelihood estimator in this role.

Briefly, the maximum likelihood estimate, $\hat{\theta}$, of an ability θ, given the item scores x_j (x_j=1 if correct, 0 in incorrect) for j=1,2,...n, is the solution of the likelihood equation

$$\sum_{j=1}^{n} \frac{x_j - G_j(\theta)}{G_j(\theta)[1-G_j(\theta)]} \cdot \frac{\partial G_j(\theta)}{\partial \theta} = 0,$$

where $G_j(\theta)$ is a general item response function and conditional independence given θ is assumed.

If this equation has a finite solution, it can usually be found efficiently by Newton-Raphson iterations, or that failing, less efficiently by direct line search. The cases where the likelihood equation does not have a finite solution are discussed below.

The limiting variance of the maximum likelihood estimator with respect to sampling of items from an infinite pool is given by the reciprocal of the test information function, as discussed above in the background material. Assuming the items carry some information about θ, it is apparent that the measurement error will be minimized if each item is selected to have maximum information at θ. The error variance of the maximum likelihood estimator will decrease as items are added and eventually the stopping criterion will be reached.

The main limitation of maximum likelihood-maximum information procedures in adaptive testing is that a finite estimate of ability does not exist when all of the examinee's responses are correct or all incorrect, or when the guessing model is used and certain unfavorable answer patterns occur (Samejima, 1973). The maximum information procedure

cannot begin, therefore, until at least one correct and one incorrect response has occurred, and with the guessing model there is a finite probability that it will fail at other times. In either case, some ad hoc rule must be adopted to keep the procedure on track. Samejima (1981), for example, has proposed a procedure for attributing an ability when responses are all correct or all incorrect. Another possibility is the "biweighted" maximum likelihood estimator proposed by Mislevy and Bock (1982). By multiplying each term in the likelihood equation by a Mosteller-Tukey (1977) biweight, one may suppress the effects of chance or other spurious responses to items early in the sequence when the provisional estimate of ability is poor. This improves the robustness of the estimator against unfavorable answer patterns.

An important feature of both major types of item selection strategies is that they continually revise the estimate of the ability, θ, at every step. Thus an estimate of ability is an inherent part of the item selection process.

Recommendations. The value and utility of item selection and scoring ultimately rests on the degree of precision and efficiency obtained. Evaluation of reliability and precision has been discussed above. Here we consider efficiency, and other ancillary issues.

IS1. The procedure for item selection and ability estimation must be documented explicitly and in detail. Essential.

IS2. The procedure should include a method of varying the items selected, to avoid using a few items exclusively. Essential.

IS3. The procedures used should include a mechanism to maintain a rough balance of correct answer options. Desirable.

Several algorithms might be used to select the next item for a candidate, conditional upon his previous responses. If the algorithm selects the most informative next item, then only the most discriminating items, that is, the items with the highest values of a_j will be selected frequently, whereas items that are very nearly but not quite as good will seldom be selected.

Whatever algorithm is used for item selection, we recommend listing the most informative items - perhaps the ten best, perhaps all whose information is at least 90% of the information one would get from the best item. Then a random selection would be made from that set of nearly equivalent items. (Or, the selection could be weighted in favor of items that had not been used as much.)

Item selection is relevant to another problem. On a standard test it is necessary to randomize, or at least mix up, the answer option that is the correct answer to the questions. It is not acceptable to have (c) be the correct answer most of the time. In the computer presentation mode, this is less of a problem because the test taker does not have the record of his past responses before him. Still, some mixing is necessary.

The effect of answer option distribution is subject to experimental study. Perhaps less able candidates favor the options encountered first (a or b). Probably there is very little chance that anyone will encounter a string of problems where the correct answer option is the same, and there may be essentially no chance at all that an examinee would notice the pattern. Recall that the examinee is getting about 1/3 of the items wrong!. Still, this is an aspect of adaptive tests that should be studied.

On a tailored test the pattern of correct options is unique to each test taker. Probably the computer system should keep track of previous correct answer options for this respondent on this test. The infrequently used options can be used to influence the choice among nearly equivalent items at each step.

Obviously, the item pool for a test should position the correct answers with equal frequency over the options. Since the sets of equally informative items from which item selection is made will be sets of items of very similar difficulty, the correct answer options should be balanced by difficulty level within each pool. Some experience will show whether special pains should be taken to keep the answer options in balance for each test-taker.

In principal, answer options could be rearranged for an item when it is presented, provided that there is no natural order for the options. But we know so little about the processes of answering items that even that slight manipulation might be dangerous.

IS4. The computer algorithm must be capable of administering designated items, and recording the response separately, without interfering with the adaptive process. Essential.

The item selection program will have to permit administering items that are being pretested or items that are being recalibrated, in the course of a regular test. The computer programs must be able to handle this possibility.

Predicting a good starting value. After the first test in the battery has been administered, additional efficiency could be gained by using a regression estimate of the examinee's ability on each subsequent test as a starting place for the tailoring process. This procedure has been used with good results by Maurelli & Weiss (1981). If this scheme is used, then the order of tests in the battery becomes important. At least, the first test should be the test having the highest correlation with all the others - the test closest to the first principal component of the battery. (That test is probably Word Knowledge, or Science Information.) The first test might well be tested to a slightly more stringent accuracy criterion (or using slightly more items) than the other tests, if it is given the added role of predicting starting values for subsequent tests.

IS5. The computer system must be able to base the choice of a first item on prior information. Essential.

The possibility has been considered of choosing a starting item for the first test on the basis of external information such as number of years of

formal schooling. We view this as unsound. First, this might be unfair to certain ethnic subgroups. Second, the test is intended to provide independent information on ability. To use ancillary information would disrupt the independence being assumed in the general prediction and counselling situation.

Stopping Rules

One of the most attractive properties of adaptive testing is the possibility of having a constant measurement error variance at all levels of ability. This not only simplifies discussion of the reliability of the test (q.v.), but it also satisfies better the assumption of homogeneity of variance in subsequent test analysis. In the regression, homogeneity of measurement error variance is a necessary condition for homogeneity of residuals. It can be attained in adaptive testing if an information criterion (or posterior variance criterion) is used to terminate the item presentations. If, for example, testing is continued until the error function is 1/10 of the population standard deviation of ability, the adaptive procedure will have a uniform reliability of $1/(1 + 1/10) = .91$, which would be acceptable in most testing applications. When the adaptive test is replacing a conventional test, the criterion should be the error variance of the latter at the ability level where it is most reliable. If the adaptive test is continued until the measurement error attains this value, the adaptive test will always be as reliable or more reliabile than the conventional test regardless of the population in which it is applied.

A simpler stopping rule is always to use the same number of items. With this procedure, the ability of some persons will be estimated more accurately than others. On the average, very low and very high ability levels will not be estimated as well as the middle levels even when the first item is selected on the basis of other relevant performance.

One disadvantage of stopping at a fixed level of measurement error variance is that some persons may need many more items than others, which may have some operational difficulties. A hybrid rule might be adopted in which testing is stopped when an acceptable level of measurement error is reached, or when a certain number of items is given, whichever happens first.

IS6. If testing continues until a specified level of measurement error variance is attained, the average number of items necessary should be reported as a function of ability level. Desirable.

IS7. If testing is stopped after a fixed number of items is given, the achieved level of measurement error variance should be reported. Essential.

This recommendation is discussed in the section on reliability.

IS8. If a hybrid rule is adopted for stopping testing, a report should be made of both the measurement error level achieved as a function of ability; Essential, and the average number of items used. Desirable.

Equating of Ability Scales

When initially implementing a CAT procedure, it may be necessary to test some individuals with the previously used P&P procedure until full implementation is achieved. It may also be desirable to compare scores on the CAT procedure to previously developed norms. In both of these cases it is necessary to form tables to convert the ability estimates from the CAT procedure to the score scale from the traditional tests. The formation of this table is called equating. If the equated ability estimates are to be interpreted properly, the accuracy of this equating must be demonstrated.

Q1. When a paper and pencil test and a computerized test on the same content are administered to the same person, the interpretation of the scores must be shown to be the same. Essential.

The evaluation of the equating of test forms is a very difficult task since there is usually no standard for comparison. Therefore, to get some indication of quality, similarity of equating results is often used. In the case of evaluating the equating of the CAT score scale to a paper and pencil test, the following procedure could be used. First, two parallel pools of test items (A and B) that have item parameters on the same scale should be created. One group of individuals would then be administered tests using both the paper and pencil form and the CAT procedure using the Pool A. Based on this administration an equivalent score table would be produced using one of the many procedures available (eg. IRT, Lord 1981b, or equipercentile, Lord, 1981a). The same process could also be followed using the paper and pencil test and Pool B. If the equating is acceptable, the ability estimates determined using Pool A should be equated to the same paper and pencil score as those obtained from Pool B.

Q2. The rank order of individuals ordered by both a CAT and paper-and-pencil instrument on the same content should be approximately the same. Essential.

Q3. The ability estimates obtained from the CAT procedure should be measuring the same trait as the scores from the paper and pencil test. Essential.

Evidence for the above two recommendations can be obtained in a manner similar to that used in the validity section of these standards. Care must be taken to compensate for possible nonlinearity of the relationship between raw scores and ability estimates. One procedure is to use expected true scores rather than ability estimates, θ, in the analysis. Another is to use scores on the equated scales.

The problem of calibrating the CAT tests needs more thorough study than we have been able to give it. As noted in the introduction, scaling errors can have important effects (Department of Defense, 1980a,b). Careful study is needed of the current calibrations of the ASVAB (Maier & Grafton, 1981 a,b).

It will be advisable to make use of a very well-established data base, the Profile of American Youth (Department of Defense, 1982) which

established national norms on ASVAB Form 8A.

Q4. Comparative data should be available on ASVAB form 8A and the CAT item pools so that CAT can be equated to the 8A normative data. Very desirable.

Human Factors

The CAT will be presented on some kind of computer display, and responses will be made on some kind of keyboard. Because this method of test administration differs markedly from the conventional P&P test, special care should be taken to insure that the devices and the environment in which the test is taken be conducive to good test performance. A variety of specific factors will be noted first, and then the cumulative effect of the novel environment will be considered.

HF1. The environment of the testing terminal should be quiet and comfortable, free of distractions. Essential.

It is an axiom of test adminstration that the environment be quiet and comfortable. This is widely understood and is relatively easy to achieve in a P&P mode. However, computer terminals are often set up in large rooms that have considerable ambient noise, and activity. It is important to stress that such an environment is inappropriate. CAT requires a quiet environment, free of distractions. A separate cubicle for each terminal would be desirable. (A mundane point is that a paper and scratch paper must be available.)

It is important to note that the test could in principal be given in a noisy, frenetic environment, so long as the same type of environment was provided for everyone. Almost always, a noisy, frenetic environment means lack of control over the environment, so that everyone is not tested under the same conditions. The important criterion is fairness. Everyone must have the same chance to succeed. Also, the quiet environment makes the test more nearly a pure test of cognitive ability, skill or knowledge. In a noisy environment, the test would also have a component of ability to work in such environments. That might be an interesting facet; if so it should be explicitly and separately evaluated.

HF2. The display screen should be placed so that it is free from glare. Essential.

Unless care is taken in the design of the display device, and the placement of consoles, the room lights, or sunlight from nearby windows could be reflected by the surface of the display screen, greatly reducing legibility and increasing testing time.

One common method of reducing the possibility of glare from overhead lights is to place the screen surface very nearly vertical. When so tilted, the screen must then be placed relatively high off the table so that it can easily be viewed by a seated person. The combination of tilt and height should be watched.

To some extent, equipment design can help reduce the chance of glare, but eventually this will be the responsibility of the operational personnel and proctors. Instructions to them must be explicit about glare as well as other aspects of the environment.

HF3. The legibility of the display should be assessed empirically. Desirable.

The legibility of the display and the speed with which it can be read are important factors. Many people are not yet accustomed to reading material

from a computer screen. The letters on the screen are less distinctive, having less detail; the contours are also necessarily less sharp. Normally the screen shows light characters on a dark ground, which is the reverse of print. In any case, it will be important to check the speed and accuracy of viewer comprehension of material on the proposed display screen, relative to ordinary print.

HF4. The response device should be carefully designed; the display screen should give a clear positive indication of the response selected; the testtaker should be able to alter his response if he thinks he pushed the wrong button. Essential.

The accuracy of response is of concern. Once the examinee has decided that (c) is the correct answer, how much difficulty does he have in indicating his choice to the computer, and in verifying that he indicated what he intended. We would expect the computer terminal to have a definite advantage here, by comparison with the usual answer sheet with bars to be blackened. The main advantage is place-keeping. In a CAT, the examinee cannot mark the wrong item, but can he quickly find the right button to press (or its equivalent, with other response devices)?

Everyone makes occasional errors. The screen must provide immediate (say within 1/2 second) feedback to the examinee about which response was actually selected. Then some mechanism must be provided to permit changing the response if it was in error. One possibility is to accept the response immediately but permit the respondent to change it, if he responds within some short time (such as three seconds.) Change could be signalled by a new response, or by pressing a separate "change" button, followed by a new response. If the next item is ready for presentation before the cancellation interval, it can either be held or its availability can automatically terminate the cancellation interval, which would then be variable. If the former, then the fed-back response could blink during the cancellation interval.

Another procedure would require the respondent to signal positively, by pressing a "verify" button, that the response fed back by the system was the response intended. This would be like the "return" key on most computer terminals. Requiring verification requires more button pushes, and makes the response process more complex, hence more prone to errors. But it may save inadvertant responses, and it does require the respondent to make sure that the recorded response was the intended response. Empirical observations are needed to guide this decision. We tend to favor the "verify" button, but the choice is by no means obvious.

HF5. The effect of the response mechanism on the speeded tests must be determined. Essential. (See later section on the speeded tests.)

Responding is especially critical in the speeded tests, in which speed of cognitive functioning is at issue. Here, responding should be especially easy and compatible with the item display format. Again, we expect the computer terminal to be superior to marking answers on an answer sheet. Note here that we are only concerned with selecting one out of 4 or 5 alternatives. The possibility of a free-answer format, especially in the numerical operations test, is intriguing. It should definitely be studied for future use.

Part of the difficulty in choosing the response mode, discussed in the previous section, is that a different mode may be needed for the speeded tests. Probably verification should not be required in the speeded tests, and correction should not be allowed. Also, the next item should appear as soon as possible. There should be a fixed interval between the response and its feedback to one item and the presentation of the next item. One second might be enough; two seconds would seem to be an upper limit. Note that no tailoring is necessary for the speeded tests.

HF6. The display must be able to include diagrams that have fine detail. Essential.

Legibility of diagrams and line drawings is an especially vexing problem. Here the limitations of the display may have to interact with the item production and selection system. Some drawings may have fine detail that is irrelevant. Others may have fine detail that is relevant, and that is obscured on the computer display screen. It would be best if TV-quality figures and diagrams could be used. The ordinary microcomputer terminal has at best about 200 lines of resolution, not enough for some of the drawings in the current versions of the ASVAB. Graphics terminals with a resolution of at least 400 lines can produce acceptable figures.

HF7. The test proctor should be able to monitor test performance and should be signalled automatically when irregularities occur. Very desirable.

The test proctor should be warned by the system if an active terminal has not produced a response in some reasonable time period. Other erratic behavior, such as excessive responses, responses within 0.1 second of item presentation, or similar peculiar patterns, may mean that the examinee does not understand how to use the terminal, or it may mean that the terminal is operating incorrectly.

HF8. The test terminals should always be in proper working order. Essential.

Any flaw in the terminal, such as sticky keys, may disrupt test performance. Even on untimed tests, persons at a computer terminal usually feel (assume) that fast response is required, and inability to do so may disconcert them. In general, a schedule of frequent regular, maintenance should be established, to keep the terminal display clean and in proper working order.

HF9. The terminal response system should include some additional buttons or similar controls for future use. Desirable.

The ordinary item on a CAT will be an item that is shown all at once on the display screen and that requires a selection of one from a few alternative responses. But the equipment should permit constructed responses, especially for numerical problems. Also, some future items may involve successive displays that may optionally be shown again at the examinee's request. Flexibility for future development is important.

Special Issues

Following are recommendations about some issues that could not be classified above.

The Speeded Tests. Two of the ASVAB subtests, Numerical Operations and Coding Speed are highly speeded tests. The current theory of adaptive testing does not apply to speeded tests. Item response theory assumes a power test, and would prefer that every respondent answer every item presented to him. Thus the speeded ASVAB tests cannot now be made adaptive.

However, the speeded tests can and should be administered by computer in the CAT environment. Here the particular design of the display and response devices will play a role in the difficulty of the test. Very likely the computer version will permit students to work faster than the paper-and-pencil version, because a keyboard response is probably faster and less prone to error than marking an answer sheet. It is hoped that the amount of difference between the computer and the paper-and-pencil versions will be constant for all test-takers, but this must be checked. Actually, a constant difference for all test takers is unlikely. Most of the reasons for a difference, such as legibility, difficulty of place keeping on the answer sheet, etc. tend to apply at the item level; there is more likely to be either a constant difference per item, or a proportional difference per item. In either case there would then be a slight change in the test score distribution for the higher scores; this is not likely to be a serious problem, but again it should be checked.

U1. The computing system must be carefully designed for the speeded tests so that the system itself adds no variability in testing time. Essential.

Each new item should be presented a fixed time after the respondent presses the response button. This fixed time should be as short as possible consistent with the requirement that it be essentially constant. An inter-item time of one second would seem an upper limit, and 500 milliseconds might be better. (As a system specification, the fixed time interval should have some tolerance level, such as $\pm$ 5%.)

U2. The calibration tables for the speeded tests must be prepared using data from the operational equipment. Essential.

The equipment will have a main effect even if it doesn't contribute to the measurement error variance. Since the score on these tests is essentially the number of items correct in a fixed time interval, the speed of reading the display and using the response mechanism is a part of the score. Thus calibration must involve the actual equipment to be used.

One obvious implication is that special scoring problems will arise if the computer display and response equipment is not identical for all test-takers. If more than one kind of equipment is introduced, separate score conversions will have to be worked out for each type of equipment. Thus whenever a new an better model of test terminal is introduced, the speed tests will have to be recalibrated. Also, norms should be developed for alternative equipment. In case of mobilization, for example, there may be a need to use standard computer terminals, somehow. This creates a

severe problem for the tests with diagrams (a booklet of diagrams might have to be provided), and it also creates a severe problem for the norms of the speeded tests. It may be that most standard terminals can be made enough alike by overlays on the keyboard, so that one or two alternative calibrations would be needed. And it might be that equipment differences are too small to require different norms, but this cannot be assumed. Empirical evidence is needed.

U3. The equipment should permit recording the time between item presentation and item response, for each item for each respondent. Highly desirable.

Eventually, we would hope that use could be made of response times to individual items. Research will be needed on this topic. Probably time to the nearest 1/60 second would be sufficient, but time to the nearest millisecond might be handy. Note that the response time itself may be as small as 0.5 second.

We have assumed here that the items will be presented serially, one at a time. This would be standard practice in a psychological laboratory. Someone has suggested that several items be displayed at once, the display changing to a new batch when all of the first set have been answered; the score would be the time to respond to a fixed set of items, or the total number of correct items in a fixed time period. Although this format might be more nearly like the P&P test, it retains some of the placekeeping nature of the P&P test, which is a procedural confound. The time wanted is only the time to do the cognitive operations. If placekeeping is to be tested, it should be done separately. Further, all the other tests will have been presented one item at a time and a screenful of items would be confusing to the test-taker.

Omits

Expert opinion differs concerning whether test-takers should be permitted to omit an item on a CAT. If the test taker doesn't understand the item, forcing a response may mean forcing a guess, which adds error. On the other hand, the item has in principal been selected as the most informative item about this person's ability. It would be unfortunate to lose the utility of this very informative item.

If omitting is permitted, should the omitted item be replaced by one of equivalent difficulty, or by a slightly easier item? Omitting means, at least in part, a failure, so a slightly easier item would seem appropriate. Merely presenting an easier item represents a slight penalty in a short tailored test.

If the respondents are permitted to omit items then the best psychometric procedure would involve the use of a graded response model. It is often found that an omit deserves more credit than an incorrect response. However, it would be very difficult to accumulate enough data through spontaneous omitting of items. Also not much experience has yet accrued concerning graded response models. Thus at present it is difficult to determine what to do with omitted responses.

U4. For the present, omits should not be permitted . Desirable.

This seems the most defensible alternative, psychometrically. However, we are not comfortable with this recommendation, and strongly urge additional research on various alternatives. Graded response models are practical with a computer, and should be explored, with and without omitting.

Item Bias

It is important that insofar as possible, the items on the test should not be offensive to any group of persons, nor should they favor any one group more than another, apart from the ability being tested. Of course, one group may actually surpass another on some ability. Men, for example, may score higher than women on shop knowledge because on the average they know more about shop. But even when such overall group differences are controlled statistically, some items may show a group difference for other, irrelevant reasons. Such items would be considered biased.

U5. All potential items for the CAT item pools should be screened in an attempt to identify and discard items that are offensive to ethnic groups, or to women, or men. Items should be screened by judges qualified to identify such biased content. (Essential)

U6. For each item pool, statistical studies should be done of item bias, by comparing subgroup performance on the items. (Highly desirable).

Several statistical procedures have been used, and no one is generally believed to be better than another. With conventional item analysis, the simplest procedure is to determine difficulty (equated delta plots) for identifiable subgroups. Separate analyses should be done comparing Whites and Blacks, and comparing men and women. With IRT, a straightforward procedure is to obtain item parameters separately for the two groups, and to compare the item response curves visually, as well as statistically. Because the groups may have large average differences, an analysis should also be done with a sample from the majority population that has the same score distribution as the minority groups, generating the IRC's from groups of comparable average ability. Methods for studying item bias have been discussed by Shepard (1981), Levine (1982), and in Berk (1982).

Because such studies have recently been made of one current form of the ASVAB, (Bock & Mislevy, 1981) the need for item bias studies is not pressing. They should be done, but can be done after more critical problems have been solved. (See also Wing, 1980.)

We note that statistical studies of item bias seldom find evidence of item bias. (See, for example, Linn et al, 1981.) Apparently, the screening process usually works quite well, so no blatantly biased items are missed. Items identified by statistical methods as possibly biased are frequently baffling, in the sense that nothing in their content seems at all likely to lead to unusual group differences. This does not mean that statistical studies should not be made, but that sometimes the results of such studies are difficult to interpret.

Such studies are also difficult for operational reasons. There may be relatively few cases in a minority group. It may be necessary to restrict the study to a subset of items, or else the study may have to wait until the requisite cases have been accumulated.

IV. Procedural Details and Future Prospects

Comments on the Procedures

A variety of special procedures must be adopted to permit a computer to conduct a testing session, as well as to store the CAT item pool in the first place. A method is needed to estimate the item parameters (a, b, c). An algorithm is needed to tailor the test, and to score each respondent. Specific decisions must be made about many details, including balancing item options, choosing a starting value for each person on each test, how to present the instructions, and management of rest periods.

P1. All procedures must be documented and described in enough detail so that the procedures could be reproduced on another computer from the documentation alone. (Essential)

The importance of explicit documentation cannot be overemphasized. Evaluating the psychometric quality of the test may depend on knowing some details of the procedure. For example, the manner of estimating item parameters is critical to the equating process. Future evaluations and research projects may require knowing certain parts of the procedure. Certainly the project must never be in the position of viewing the computer as a mysterious black box with a mind of its own. It should never be necessary to attribute a result to some inexplicable decision of the computer.

The details of the administration of the CAT are not in the province of this committee, except as they affect the evaluation. Still, the committee wishes to provide some comments on the procedures that should be considered by those in charge of administration.

1. Rest periods. The current ASVAB requires about 2 1/2 hours plus rest breaks and instructions. Presumably the CAT version will reduce this to about 1 1/4 to 1 1/2 hours. Test-takers will have to be given a short rest break, at least once or twice in the total testing session. Continuous interaction with a computer display can be very tiring. The system must be programmed to provide these breaks, and to respond to some kind of signal that the break is finished. The system must also be able to accommodate unscheduled breaks, in emergency situations.

Almost certainly, some test-takers will complain of eye strain as a result of watching the display screen more or less constantly for about 90 minutes or more. Probably the eye strain is an excuse for the more fundamental problem of cognitive strain. The test will seem moderately difficult to everyone, since ideally they will only get about two-thirds of the items correct. Perhaps some initial warning to that effect could be provided. Still, one or two breaks between tests would be advisable.

2. Instructions and sample items. The mode of presenting and checking instructions for each test deserves careful attention. What if the examinee answers a sample item incorrectly? Probably he should be forced to continue to respond until he gets it correct. But should an easier sample item then be displayed, the process continuing until he can answer

one of the sample items correctly on the first try? Should the test supervisor be called to check the terminal? Perhaps the examinee doesn't understand the use of the response buttons. These issues deserve careful attention and planning.

Should the examinee be permitted to review the instructions, once he has started on the tests? Probably so, but this may be subject to revision if the examinees overuse the options. This creates the problem of designing a means for the examinee to review the instructions and to cancel the review.

The examinee will not be permitted to return to a previous item once the next item has been presented. In the P&P version, of course, the examinee may erase, go back, and reconsider ad lib, probably to his detriment.

3. Response time of terminal. The computer should respond within one second to test-taker responses and key presses. A few seconds may be tolerable between successive items on unspeeded tests, though this time should be kept as short as possible. A design objective of two seconds should be established. Three seconds between items might be tolerable, but five seconds would seem unreasonable. Shorter response times are needed for the speeded tests.

4. Monitoring quality. A maximum elapsed time should be established between the presentation of an item to the person and the receipt of a response from the person. This maximum interval may be one minute, or may be 30 seconds - it should certainly be an adjustable system parameter. Experience will dictate the best setting; this default is necessary to guard against a test taker who is not alert, or a terminal that is not working properly.

Research and Development

The introduction of a computerized adaptive version of the ASVAB will represent the first large-scale use of CAT. The recommendations in this report are designed to insure that CAT provides the expected substantial increase in efficiency with no loss in quality. The procedure will need to be monitored to insure that it is operating in the ways that are expected. Also, the present efforts have mainly been toward establishing an operational test. Maintenance of the testing procedure on an operational basis will require added attention. New items will have to be added from time to time, old items will have to be retired. The maintenance and evaluation of the new version of the test will require extended attention from research and development specialists.

It will be necessary to check and refine the psychometric procedures employed in CAT. Present technology has been developed in the absence of extensive experience with CAT. As experience accrues, procedures must certainly be monitored, and may well require some modification. A number of theoretical and practical questions remain to be answered. For example, how sensitive is the process to biased estimates of item parameters? Is the item selection algorithm working as expected? If omitting is permitted, is it widespread? Is there evidence of inappropriate responses (low scoring candidates getting difficult items correct, or high scoring candidates missing easy items.) Is the equating and norming of scores satisfactory? Statistical methods are needed for assessing the unidimensionality of item banks. In general, statistical methods are needed for analyzing the kind of item response data that emerges from the CAT.

Throughout this report we have identified other issues that require experimental study. Many choices have had to be based on judgment, rather than evidence. Studies should be done on all these issues, so that knowledge can replace opinion, to provide a solid basis for CAT procedures. Research is needed in every area - dimensionality, reliability, validity, item parameter estimation and linking item pool characteristics, item selection ability estimation, and scale calibration. An extensive program of psychometric research is a necessary adjunct of a CAT system.

In addition to these technical matters, there are many opportunities for CAT to make fundamental improvements in the personnel selection and classification system. The introduction of the computer-administered adaptive version of the ASVAB has great potential for improved personnel assessment in the Armed Forces, quite a part from the immediate savings realized in the recruit testing process. To realize this potential, further research and development projects are needed to develop the most promising possibilities.

The most likely benefits from CAT are improved measurements of abilities now included in the ASVAB and the addition or change of abilities now measured. The present ASVAB is well designed, but the tests in the

battery are intercorrelated to an extent that detracts from potential validity. That is, the present tests are not sufficiently distinct. Validity is likely to be improved if tests measuring different aspects of ability are included. Validity of the ASVAB for predicting performance in various technical schools is adequate, but modest. There is much room for improvement. Any improvement in validity translates directly to economic benefits in reducing dropout and failure rates. Both the services and the individual recruit lose if the recruit is placed in a specialty for which he or she is not qualified.

Various possibilities can be explored for obtaining more information from the present test. Additional measures, such as reaction time, can be taken (Micko, 1969, Thissen, 1980.) The items can be presented in a format that requires the recruit to continue choosing answer options until he hits the correct answer. Free answer formats can be tried, at least on the test of arithmetic operations. It is by no means obvious how to use this additional information, so considerable exploration will be required.

A better way to get more information is to use new item types. One is already being studied by McBride (1980). He is altering the presentation of reading comprehension items so that the passage appears first. Then, when the respondent is ready, the passage disappears and is replaced by the question about the passage. This makes the test sufficiently different from the current version that it would not be wise to include in the original CAT version of the ASVAB. It has a memory component that may make the test more distinctive, and hence not strictly comparable with the parallel P&P mode. It is however a goal, but for the future since it may be more valid for many uses.

The best way to get more information is to test additional aptitudes and skills. Other tests or test items include spatial visualization, not now a part of ASVAB, items with moving parts for mechanical ability tests, judgments about collision of two or more moving elements, discrimination of temporal intervals, and more. The list is endless. Although some of these possibilities will not prove to be useful, others surely will be sufficiently promising that they would considerably improve the predictability of recruit performance.

The relationship and possible contribution of computerized adaptive testing to the process of counselling and placement procedures needs study. Instead of asking, "Will this particular recruit pass School A? School B?", the Armed Services will increasingly be asking, "How can we best use this recruit?" (Or from the recruit's perspective, "How can this recruit best realize his or her potential?") This implies using measures like the ASVAB for placement, rather than selection. Over the years, the Armed Services have studied the placement problem, but there has been little attempt to design ASVAB with an eye on placement rather than selection. In a CAT environment, in addition to new tests, there are other possibilities. Possibly everyone need not be tested on all variables. The counselling opportunities can be explored in the context of existing work on placement systems in the Air Force (Hendrix, Ward, Pina, & Harvey, 1979), and the Navy (Horst & Sorensen, 1976).

The many possibilities for research and development represent opportunities to realize additional economic benefit from a CAT system. The operational benefits of adopting CAT are manifest; the possibilities mentioned here are ways of getting added value from the investment.

Reference Notes

1. Levine, M.V. Identifying different item response curves. Prepublication draft, 1982.

2. Ree, M.J. An automated classification system. Paper presented at the American Psychological Association Convention, September 1979.

3. Reiser, M. MAP estimation of item parameters. Presented at Psychometric Society meetings, May 1981.

4. Stout, W. *Evaluating local independence.* Talk given at ONR contractor's meeting, Memphis, Tenn., October, 1981.

5. Swaminathan, H. Bayes estimation of item parameters. Presented at Psychometric Society Meetings, May 1981.

References

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association. 1974.

Andersen, E.B. Conditional inference for multiple-choice questionnaires. British Journal of Mathematical and Statistical Psychology, 1973, 26, 31-44.

Andersen, E.B. Discrete statistical models with social science applications. Amsterdam: North-Holland, 1980.

Andersen, E.B. and Madsen, M. Estimating parameters of the latent population distribution. Psychometrika, 1977, 42, 357-374.

ASVAB Working Group. History of the Armed Services Vocational Aptitude Battery (ASVAB). Washington, D.C.: Office of the Assistant Secretary of Defense, (Manpower Reserve Affairs and Logistics.) March, 1980.

Bartholomew, D.J. Factor analysis for categorical data (with Discussion). Journal of the Royal Statistical Society, Series B, 1980, 42.

Berk, R.A. (Ed.) Handbook of methods for detecting item bias. Baltimore, Md.: The Johns Hopkins University Press, 1982.

Binet, A. Les idees modernes sur les enfants. Paris: Ernest Flamorion, 1909.

Birnbaum, A. On the estimation of mental ability. Series Report No. 15. Project No. 7755-23, USAF School of Aviation Medicine Randolph Air Force Base, Texas, 1958.

Birnbaum, A. Some latent trait models and their uses in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. Statistical theories of Mental Test Scores. Reading, Mass.: Addison, Wesley, 1968.

Bock, R.D. & Aitkin, M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 1981, 46, 443-459.

Bock, R.D., and Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Bock, R.D. & Mislevy, R.J. Data quality analysis of the Armed Services Vocational Aptitude Battery. Chicago: National Opinion Research Center, August, 1981.

Box, G.E.P. Problems in the analysis of growth and wear curves. Biometrics, 1950, 6, 362-389.

Carroll, J.B. The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 1946, 10, 1-19.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-32.

Department of Defense. Aptitude testing of recruits. A report to the House Committee on Armed Services. Washington, D.C.: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics), July 1980a.

Department of Defense. Armed Services Vocational Aptitude Battery (ASVAB) Information Pamphlet. U.S. Department of Defense, 1 October 1980b.

Department of Defense. Profile of american youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics.) Washington, D.C., March 1982.

Fischer, G.H. Linear logistic test model as an instrument in educational research. Acta Psychologica. 1973, 37, 359-374.

Fischer, G.H. Individual testing on the basis of the dichotomous Rasch model. In L.J.T. Van der Kamp, W.F. Langerak, and D.M.N. de Gruijter (Eds). Psychometrics for educational debates. Chichester: Wiley, 1980.

Fischl, M.A., Ross, R.M., McBride, J.R., Valentine, L.D., Mathews, J.J., Swanson, L., Wiskoff, M.F., & Wilfong, H.D. Validity of the Armed Services Vocational Aptitude Battery for predicting performance in service technical training schools. Technical Research Report 77-4, Directorate of Testing, U.S. Military Enlistment Processing Command, Fort Sheridan, Il., March, 1978.

Friedman, D., Steinberg, A., & Ree, M.J. Adaptive testing without a computer, AFHRL-TR-80-66. Air Force Human Resources Laboratory, Brooks Air Force Base, TX, March 1981.

Goldstein, H. Dimensionality, bias, independence, and measurement scale problems in latent trait test score models. British Journal of Statistics and Mathematical Psychology. 1980, 33, 234-246.

Goodman, L.A. Analyzing qualitative categorical data: log-linear models and latent structure analysis. Cambridge (Mass.): Abt Associates, 1978.

Haberman, S.J. Maximum likelihood estimates in exponential response models. Annals of Statistics, 1977, 5, 815-841.

Hambleton, R.K. & Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement. 1977, 38, 75-96.

Harman, H., Helm, C.E., & Loye, (Eds.) Computer-assisted testing. Princeton, N.J.: Educational Testing Service, 1968.

Hendrix, W.H., Ward, J.H., Jr., Pince, M., Jr., & Harvey, D.L. Pre-enlistment person-job match system. AFHRL-TR-79-29, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas, September, 1979.

Holland, P.W. When are item response models consistent with observed data? Psychometrika, 1981, 46, 79-92.

Holtzman, W.H., (Ed.) Computer-assisted Instruction, Testing, and Guidance. New York: Harper & Row, 1970.

Horst, P., & Sorenson, R.C. Matrix transformation for optimal personnel assignments. Technical Note 77-5, Navy Personnel & Research and Development Center, San Diego, CA, December 1976.

Jaeger, R.M., Linn, R.L., & Novick, M.R. A review and analysis of score calibration for the Armed Service Vocational Aptitude Battery. Washington, D.C.: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics), June 1980.

Johnson, P.O. & Neyman, J. Tests of certain linear hypotheses and their applications to some educational problems. Statistical Research Memoirs, 1936, 1, 57-93.

Jones, D.H. On the adequacy of latent-trait models. Program Statistics Research Technical Report No. 80-8. Educational Testing Service, Princeton, N.J., April, 1980.

Joreskog, K.G. & Sorbom, D. LISREL: analysis of linear structural relationships byu the method of maximum likelihood, Chicago, Illinois: National Educational Resources, Inc., Inc., 1978.

Kingsbury, G.G. & Weiss, D.J. A validity comparison of adaptive and conventional strategies for mastery testing. Research Report 81-3, Psychometrics Methods Program, Dept. of Psychology, University of Minnesota, MN, September, 1981.

Koch, W.R. & Reckase, M.D. A live tailored testing comparison study of the one- and three-parameter logistic models. Research Report 78-1. Tailored Testing Research Laboratory, Educational Psychology Department, University of Missouri, Columbia, MO, June, 1978.

Kreitzberg, C.B. & Jones, D.H. An empirical study of the Broad-Range Tailored Test of Verbal Ability. RR-80-5. Educational Testing Service, Princeton, N.J., May, 1980.

Laird, N.M. Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association, 1978, 73, 805-811.

Lawley, D.N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh. Series A, 1943, 61, 273-287.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.C. Item bias in a test of reading comprehension. Applied Psychological Measurement. 1981, 5, 159-173.

Linn, R.L., Rock, D.A., & Cleary, T.A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 19, 124-146.

Lord, F. A theory of test scores. Psychometrika Monograph #7. 1952.

Lord, F.M. An analysis of the verbal scholastic aptitude test using Birnbaum's three parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

Lord, F.M. The standard error of equipercentile equating. RR-81-48, Educational Testing Service, Princeton, N.J., Nov., 1981a.

Lord, F.M. Standard error of an equating by item response theory. RR-81-49, Educational Testing Service, Princeton, N.J., Nov.,

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Maier, M.H. & Grafton, F.C. Aptitude composites for ASVAB 8, 9, and 10. U.S. Army Research Institute, Alexandria, VA, May, 1981b.

Maier, M.H. & Grafton, F.C. Scaling Armed Services Vocational Aptitude Battery (ASVAB) Form 8AX. Research Report 1301, U.S. Army Research Institute, Alexandria, VA, January, 1981a.

Maurelli, V.A. & Weiss, D.J. Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries. Research Report 81-4 Psychometric Methods Program, Dept. of Psychology, Univ. of Minnesota, Minneapolis, MN, Nov., 1981.

McBride, J.R. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1977, 1, 121-140.

McBride, J.R. Adaptive verbal ability testing in a military setting. In Weiss, P. (ed.) Proceedings of the 1979 Computerized Adaptive Testing Conference. Dept. of Psychology, Univ. of Minnesota, Minneapolis, Minn., Sept., 1980.

McKinley, R.L. & Reckase, M.D. A comparison of the ANCILLES and LOGIST parameter estimation procedures for the three-parameter logistic model using Goodness of fit as a criterion. Research Report 80-2, Tailored Testing Research Laboratory, Educ. Psych. Dept., Univ. of Missouri, Columbia, MO., Dec., 1980.

McKinley, R.L. & Reckase, M.D. A comparison of a Bayesian and a maximum likelihood tailored testing procedure. Research Report 81-2. Tailored Testing Research Laboratory, Educational Psych. Dept., Univ. of Missouri, Columbia, MO, June, 1981a.

McKinley, R.L. & Reckase, M.D. A comparison of procedures for constructing large item pools. Research 81-3, Tailored Testing Research Laboratory, Educ. Psych. Dept. Univ. of Missouri, Columbia, MO, August, 1981b.

Micko, H.C. A psychological scale for reaction time measurement. Acta Psychologica, 1969, 30, 324-

Mislevy, R.J., & Bock, R.D. Biweight estimates of latent ability. Educational and Psychological Measurement, 1982, 42 (in press).

Mosteller, F. & Tukey, J. Exploratory data analysis and regression. Reading (Mass.): Addison-Wesley, 1977.

Muthen, B. Contributions to factor analysis of dichotomized variables. Psychometrika, 1978, 43, 551-560.

Owen, R.J. A Bayesian approach to tailored testing. Research Bulletin 69-92, Educational Testing Service, Princeton, N.J. 1969.

Owen, R.J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Patience, W.M. & Reckase, M.D. Operational characteristics of a one-parameter tailored testing procedure, Research Report 79-2, Tailored Testing Research Laboratory, Educ. Psych. Dept., Univ. of Missouri, Columbia, MO, October, 1979.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark. Dansmarks Paedogogiske Institute, 1960.

Reckase, M.D. Ability estimation and item calibration using the one- and three-parameter logistic models: A comparative study. Catalog of Selected Documents in Psychology, 1978, 8, 71. Ms. 1737.

Reckase, M.D. The formation of homogenous item sets when guessing in a factor in item response. Research Report 81-5, August, 1981.

Ree, M.J. Estimating item characteristics curves. Applied Psychological Measurement. 1979, 3, 371-385.

Ree, M.J. The effects of item calibrations sample size and item pool size on adaptive testing. Applied Psychological Measurement, 1981, 5, 11-19.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph Supplement, No. 17, 1969.

Samejima, F. A comment on Birnbaums's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38, 221-233.

Samejima, F. The use of the information function in tailored testing. Applied Psychological Measurement. 1977a, 1, 233-247.

Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika, 1977b, 42, 163-191.

Sampson, A. Stepwise BAN estimators for exponential families with multivariate normal applications. Journal of Multivariate Analysis, 1976, 6, 167-175.

Shepard, L.A. Identifying bias in test items. In Green, B.F. (ed) Issues in Testing: Coaching Disclosure and Ethnic Bias. New Directions for Testing and Measurement. San Francisco. Jossey-Bass, 1981.

Sympson, J.B., Weiss, D.J., & Ree, M. Predictive validity of conventional and adaptive tests in an Air Force Training Environment. AF HRL-TR-81-40. Air Forces Human Resources Laboratory, Brooks Air Force Base, TX., March, 1982.

Thissen, D. Latent trait scoring of timed ability tests. In Weiss, D.J. (ed) Proceedings of the 1979 computerized adaptive testing conferences.

Tucker, L.R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.

Urry, V.W. Tailored Testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Urry, V.W. Tailored testing, its theory and practice. Part II. ability and item parameter estimation, multiple ability application, and allied procedures. NPRDC TR81 Navy Personnel Research & Development Center, San Diego, CA. Nov. 1981.

Urry, V.W. & Dorans, N.J. Tailored testing, its theory and practice. Part I: The basic model, the normal ogive submodels, and the tailored testing algorithms. NPDC TR50, Navy Personnel Research & Development Center, San Diego, Nov. 1980.

U.S. Civil Service Commission. Proceedings of the first conference on computerized adaptive testing. U.S. Civil Service Commission Professional Series 75-6, March 1976.

Vale, C.D., Maurelli, V.A., Gralluca, K.A., Weiss, D.J., & Ree, M.J. Methods for linking item parameters. AFHRL-TR-81-10. Air Force Human Resources Laboratory, Brooks Air Force Base, TX, August, 1981.

Warm, T.A. A primer of item response theory. Technical Report 940279, U.S. Coast Guard Institute, Oklahoma City, Oklahoma, December, 1978.

Weiss, D.J. Strategies of adaptive ability measurement. Res. Rep. 74-5. Psychometric Methods Program, Dept. of Psychology, Univ. of Minnesota, Minneapolis, MN, 1974.

Weiss, D.J. Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement. Research Report 79-6, Psychometric Methods Program, Dept. of Psychology, Univ. of Minnesota, Minn. MN., 1979.

Weiss, D.J., (Ed.) Proceedings of the 1977 computerized adaptive testing conference. Dept. of Psychology, Univ. of Minn., 1978.

Weiss, D.J., (Ed.) Proceedings of the 1979 computerized adaptive testing conference. Dept. of Psychology, Univ. of Minnesota, 1980.

Wilfong, H.D. ASVAB: Technical supplement to the counselor's guide. Fort Sheridan, Illinois: Directorate of Testing, United States Military Enlistment Processing Command, April, 1980.

Wing, H. Profiles of cognitive ability of different racial, ethnic, and sex groups on a multiple abilities test battery. Journal of Applied Psychology, 1980, 3, 289-298.

Wood, R. Computerized adaptive sequential testing. (Unpublished Ph.D. dissertation, Dept. of Education, Univ. of Chicago, 1971).

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.

Wood, R.L. Wingersky, and Lord, F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Princeton (N.J.): Educational Testing Service, 1976.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 38, 97-116.

Wright, B.D., and Panchapakesan, N.A. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Yen, W.M. Using simulation results to choos a latent trait model. Applied Psychological Measurement, 1981, 5(2), 345-262.

Navy

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Chief of Naval Education and Training
Liason Office
Air Force Human Resource Laboratory
Flying Training Division
WILLIAMS AFB, AZ 85224

1 COMNAVMILPERSCOM (N-6C)
Dept. of Navy
Washington, DC 20370

1 CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217

1 Deputy Assistant Secretary of the Navy
(Manpower)
Office of the Assistant Secretary of
the Navy (Manpower, Reserve Affairs,
and Logistics)
Washington, DC 20350

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Richard Gibson
Bureau of medicine and surgery
Code 3C13
Navy Department
Washington, DC 20372

Navy

1 Dr. Patrick R. Harrison
Psychology Course Director
LEADERSHIP & LAW DEPT. (7b)
DIV. OF PROFESSIONAL DEVELOPMMENT
U.S. NAVAL ACADEMY
ANNAPOLIS, MD 21402

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508

1 Dr. Kneale Marshall
Scientific Advisor to DCNO(MPT)
OP01T
Washington DC 20370

1 CAPT Richard L. Martin, USN
Prospective Commanding Officer
USS Carl Vinson (CVN-70)
Newport News Shipbuilding and Drydock Co
Newport News, VA 23607

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 LCDR W. Moroney
Code 55MP
Naval Postgraduate School
Monterey, CA 93940

1 Commanding Officer
U.S. Naval Amphibious School
Coronado, CA 92155

1 Mr. William Nordbrock
Instructional Program Development
Bldg. 90
NET-PDCD
Great Lakes Naval Training Center,
IL 60088

Navy

1 Ted M. I. Yellen
Technical Information Office, Code 201
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

1 Technical Director
Navy Personnel R&D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Psychologist
ONR Branch Office
Bldg 114, Section D
666 Summer Street
Boston, MA 02210

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

1 Psychological Sciences Division
Code 450
Office of Naval Research
Arlington, VA 22217

1 Organizational Effectiveness
Research Programs, Code 452
Office of Naval Research
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

Navy

1 Special Asst. for Education and
Training (OP-01E)
Rm. 2705 Arlington Annex
Washington, DC 20370

1 Office of the Chief of Naval Operations
Research Development & Studies Branch
(OP-115)
Washington, DC 20350

1 Head, Manpower Training and Reserves
Section (Op-964D)
Room 4A478, The Pentagon
Washington, DC 20350

1 LT Frank C. Petho, MSC, USN (Ph.D)
Selection and Training Research Division
Human Performance Sciences Dept.
Naval Aerospace Medical Research Laborat
Pensacola, FL 32508

1 The Principal Deputy Assistant
Secretary of the Navy (MRA&L)
4E780, The Pentagon
Washington, DC 20350

1 Director, Research & Analysis Division
Plans and Policy Department
Navy Recruiting Command
4015 Wilson Boulevard
Arlington, VA 22203

1 Dr. Bernard Rimland (03B)
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Worth Scanland, Director
Research, Development, Test & Evaluation
N-5
Naval Education and Training Command
NAS, Pensacola, FL 32508

1 Dr. Robert G. Smith
Office of Chief of Naval Operations
OP-987H
Washington, DC 20350

Navy

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group (TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Ronald Weitzman
Code 54 WZ
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Douglas Wetzel
Code 310
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Robert Wisher
Code 309
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

1 Mr John H. Wolfe
Code P310
U. S. Navy Personnel Research and Development Center
San Diego, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Mr. J. Barber
HQS, Department of the Army
DAPE-ZBR
Washington, DC 20310

1 Dr. Myron Fischl
U.S. Army Research Institute for the Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Michael Kaplan
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Milton S. Katz
Training Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
Attn: PERI-OK
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 LTC Michael Plummer
Chief, Leadership & Organizational Effectiveness Division
Office of the Deputy Chief of Staff for Personnel
Dept. of the Army
Pentagon, Washington DC 20301

1 DR. JAMES L. RANEY
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

Army

1 Mr. Robert Ross
U.S. Army Research Institute for the Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Robert Sasmor
U. S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commandant
US Army Institute of Administration
Attn: Dr. Sherrill
FT Benjamin Harrison, IN 46256

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

1 U.S. Air Force Office of Scientific Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, DC 20332

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Dr. Genevieve Haddad
Program Manager
Life Sciences Directorate
AFOSR
Bolling AFB, DC 20332

1 Research and Measurment Division
Research Branch, AFMPC/MPCYPR
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

1 Dr. Marty Rockway
Technical Director
AFHRL(OT)
Williams AFB, AZ 58224

1 Dr. Frank Schufletowski
U.S. Air Force
ATC/XPTD
Randolph AFB, TX 78148

Marines

1 H. William Greenup
Education Advisor (E031)
Education Center, MCDEC
Quantico, VA 22134

1 Major Howard Langdon
Headquarters, Marine Corps
OTTI 31
Arlington Annex
Columbia Pike at Arlington Ridge Rd.
Arlington, VA 20380

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 Headquarters, U. S. Marine Corps
Code MPI-20
Washington, DC 20380

1 Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

1 Major Michael L. Patrow, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

CoastGuard

1 Chief, Psychological Reserch Branch
U. S. Coast Guard (G-P-1/2/TP42)
Washington, DC 20593

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Other DoD

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

1 Dr. William Graham
Testing Directorate
MEPCOM/MEPCT-P
Ft. Sheridan, IL 60037

1 Director, Research and Data
OASD(MRA&L)
3B919, The Pentagon
Washington, DC 20301

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defense
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 Dr. Wayne Sellman
Office of the Assistant Secretary
of Defense (MRA & L)
2B269 The Pentagon
Washington, DC 20301

1 DARPA
1400 Wilson Blvd.
Arlington, VA 22209

Civil Govt

1 Dr. Lorraine D. Eyde
Personnel R&D Center
Office of Personnel Management of USA
1900 EStreet NW
Washington, D.C. 20415

1 Jerry Lehnus
REGIONAL PSYCHOLOGIST
U.S. Office of Personnel Management
230 S. DEARBORN STREET
CHICAGO, IL 60604

1 Mr. Richard McKillip
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 William J. McLaurin
66610 Howie Court
Camp Springs, MD 20031

1 Dr. Andrew R. Molnar
Science Education Dev.
and Research
National Science Foundation
Washington, DC 20550

1 Dr. H. Wallace Sinaiko
Program Director
Manpower Research and Advisory Services
Smithsonian Institution
801 North Pitt Street
Alexandria, VA 22314

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. James Algina
University of Florida
Gainesville, FL 32611

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600, Australia

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Capt. J. Jean Belanger
Training Development Division
Canadian Forces Training System
CFTSHQ, CFB Trenton
Astra, Ontario KOK 1B0

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69978
Israel

1 Dr. Werner Birke
DezWPs im Streitkraefteamt
Postfach 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Dr. R. Darrel Bock
Department of Education
University of Chicago
Chicago, IL 60637

1 Liaison Scientists
Office of Naval Research,
Branch Office , London
Box 39 FPO New York 09510

Non Govt

1 Dr. Lyle Bourne
Department of Psychology
University of Colorado
Boulder, CO 80309

1 Col Ray Bowles
800 N. Quincy St.
Room 804
Arlington, VA 22217

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52240

1 DR. C. VICTOR BUNDERSON
WICAT INC.
UNIVERSITY PLAZA, SUITE 10
1160 SO. STATE ST.
OREM, UT 84057

1 Dr. John B. Carroll
Psychometric Lab
Univ. of No. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

Non Govt

1 Dr. Deborah Coates
Catholic University
620 Michigan Ave. NE
Washington, DC 20064

1 Dr. William E. Coffman
Director, Iowa Testing Programs
334 Lindquist Center
University of Iowa
Iowa City, IA 52242

1 Dr. Meredith P. Crawford
American Psychological Association
1200 17th Street, N.W.
Washington, DC 20036

1 Dr. Hans Crombag
Education Research Center
University of Leyden
Boerhaavelaan 2
2334 EN Leyden
The NETHERLANDS

1 Director
Behavioural Sciences Division
Defence & Civil Institute of
Environmental Medicine
Post Office Box 2000
Downsview, Ontario M3M 3B9
CANADA

1 Dr.,Fritz Drasgow
Yale School of Organization and Manageme
Yale University
Box 1A
New Haven, CT 06520

1 Dr. Mavin D. Dunnette
Personnel Decisions Research Institute
2415 Foshay Tower
821 Marguette Avenue
Mineapolis, MN 55402

1 Mike Durmeyer
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

Non Govt

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. A. J. Eschenbrenner
Dept. E422, Bldg. 81
McDonnell Douglas Astronautics Co.
P.O.Box 516
St. Louis, MO 63166

1 Dr. Benjamin A. Fairbank, Jr.
McFann-Gray & Associates, Inc.
5825 Callaghan
Suite 225
San Antonio, Texas 78228

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

Non Govt

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Frank E. Gomer
McDonnell Douglas Astronautics Co.
P. O. Box 516
St. Louis, MO 63166

1 Dr. Daniel Gopher
Industrial & Management Engineering
Technion-Israel Institute of Technology
Haifa
ISRAEL

1 Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

1 Dr. Ron Hambleton
School of Education
University of Massechusetts
Amherst, MA 01002

1 Dr. Delwyn Harnisch
University of Illinois
242b Education
Urbana, IL 61801

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

Non Govt

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Jack Hunter
2122 Coolidge St.
Lansing, MI 48906

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Professor John A. Keats
University of Newcastle
AUSTRALIA 2308

1 Mr. Jeff Kelety
Department of Instructional Technology
University of Southern California
Los Angeles, CA 92007

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
Netherlands

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

Non Govt

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. James Lumsden
Department of Psychology
University of Western Australia
Nedlands W.A. 6009
AUSTRALIA

1 Mr. Merl Malehorn
Dept. of Navy
Chief of Naval Operations
OP-113
Washington, DC 20350

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08450

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77004

1 Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

1 Professor Jason Millman
Department of Education
Stone Hall
Cornell University
Ithaca, NY 14853

1 Bill Nordbrock
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

Non Govt

1 Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army Navy Drive
Arlington, VA 22202

1 Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Cirle, NW
Washington, DC 20036

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

1 MR. LUIGI PETRULLO
2431 N. EDGEWOOD STREET
ARLINGTON, VA 22207

1 Dr. Richard A. Pollak
Director, Special Projects
Minnesota Educational Computing Consorti
2520 Broadway Drive
St. Paul,MN 55113

1 DR. DIANE M. RAMSEY-KLEE
R-K RESEARCH & SYSTEM DESIGN
3947 RIDGEMONT DRIVE
MALIBU, CA 90265

1 MINRAT M. L. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 1328
D-53 BONN 1, GERMANY

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
4 Hill Hall
Columbia, MO 65211

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

Non Govt

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 Frank L. Schmidt
Department of Psychology
Bldg. GG
George Washington University
Washington, DC 20052

1 Dr. Kazuo Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 980
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
University of Central Florida
Orlando, FL 32816

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 Dr. Thomas G. Sticht
Director, Basic Skills Division
HUMRRO
300 N. Washington Street
Alexandria,VA 22314

Non Govt

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Brad Sympson
Psychometric Research Group
Educational Testing Service
Princeton, NJ 08541

1 Dr. Kikumi Tatsuoka
Computer Based Education Research Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Douglas Towne
Univ. of So. California
Behavioral Technology Labs
1845 S. Elena Ave.
Redondo Beach, CA 90277

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

Non Govt

1 Dr. David Vale
Assessment Systems Corporation
2395 University Avenue
Suite 306
St. Paul, MN 55114

1 Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

1 Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 DR. GERSHON WELTMAN
PERCEPTRONICS INC.
6271 VARIEL AVE.
WOODLAND HILLS, CA 91367

1 Wolfgang Wildgrube
Streitkraefteamt
Box 20 50 03
D-5300 Bonn 2